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

Yang Zhong
Department of Computer Science
University of Pittsburgh
yaz118@pitt.edu

Diane Litman
Department of Computer Science and LRDC
University of Pittsburgh
dlitman@pitt.edu

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 natural language inference models. Our approach shows improved performance on top of different model baselines over several evaluation benchmarks, covering rich domains of texts, focusing on long document summarization. This underscores the significance of incorporating discourse features in developing models for scoring summaries for 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., 2024b; Cao and Wang, 2021; Kryscinski et al., 2020; Goyal and Durrett, 2021; Cao et al., 2022). 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 (QA) (Scialom et al., 2021; Fabbri et al., 2022) models. 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 methods mentioned above perform well with short texts, they struggle with longer documents (Schuster et al., 2022). Recent NLI work addresses this by selecting the input and breaking down the summary. Lengthy summaries are split into individual sentences or more minor atomic 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 a summary-level label (Zha et al., 2023; Zhang et al., 2024a; 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 the lens of discourse analysis. By analyzing the structure (here we use Rhetorical Structure Theory (RST) (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

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2050–2073
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

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, LONGEVAL, and a non-scientific domain dataset LEGALSUMM with a focus on long document summarization. Our proposed approach obtained a performance gain on multiple tasks.[1]

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 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. (2024a) 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., 2024a) or **hard aggregation** with the minimum score (Schuster et al., 2022; Yang et al., 2024). However, these approaches were primarily validated on older benchmarks, consisting of shorter texts (a few hundred input words and summaries of 2-3 sentences). There lacks a 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-inspired reweighting algorithm to calibrate the scores.*

Discourse-assisted Text Summarization Discourse factors have been known 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.*

[1] Our models and model outputs are publicly available at https://github.com/cs329yangzhong/discourse-driven-summary-factuality-evaluation

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
LegalSumm	Legal Opinions (Elaraby et al., 2023)	50	2873.87	115.64	8.36	208.28

Table 1: Summary-level task statistics on AggreFact-FtSOTA, DiverSumm, LongSciVerify, LongEval and LegalSUMM. 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).

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., 2024a), which come with annotated sentence-level errors labels. Following Zhang et al. (2024a), we denote it as DiverSumm-sent. It covers 293 document-summary pairs of which 3138 summary sentences have sentence-level annotations.[2]

Summary-level Factuality Detection Datasets We test on the AggreFact-FtSOTA split (Tang et al., 2023), DiverSumm (Zhang et al., 2024a), LongSciVerify and LongEval from Bishop et al. (2024). We additionally collect LegalSumm, a legal summarization dataset, which covers model-generated summaries from the CanLII (Canadian Legal Information Institute) dataset (Xu et al., 2021; Elaraby et al., 2023) with document-level factuality labels annotated by legal experts.[3] 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.

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), AI-generated text detection (Kim et al., 2024b), and summarization (Marcu, 1998; Xu et al., 2020). **RST** predicts tree structures on the grounds of underlying coherence relations that are 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 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).

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 summa-

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

[3] We provide details of the dataset in Appendix B.

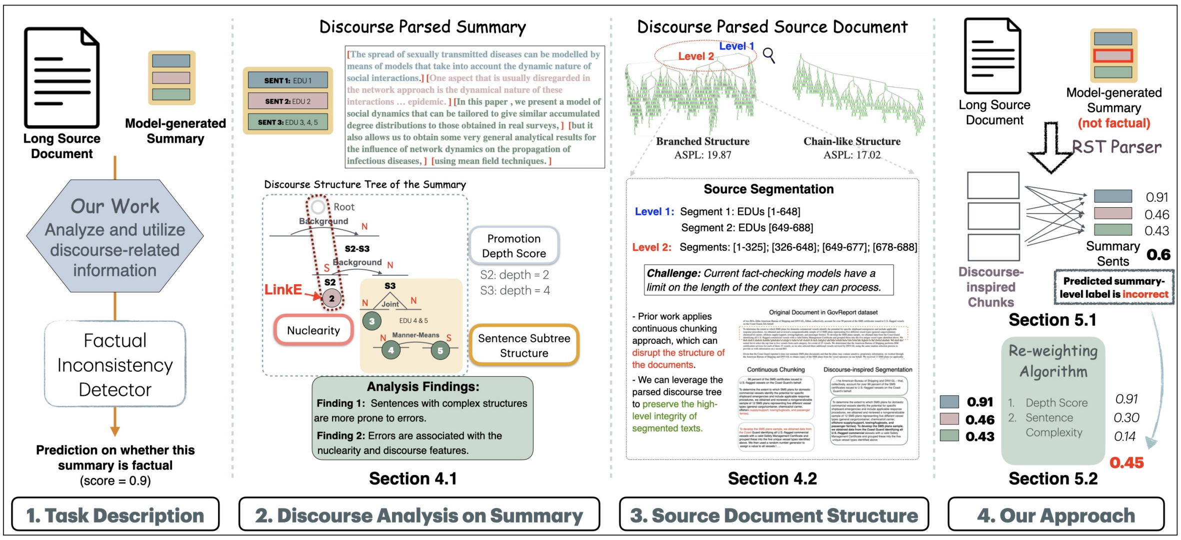

Long Source Document
Model-generated Summary
Our Work
Analyze and utilize discourse-related information
Factual Inconsistency Detector
Prediction on whether this summary is factual
(score = 0.9)
1. Task Description
Discourse Parsed Summary
SENT 1: EDU 1
SENT 2: EDU 2
SENT 3: EDU 3, 4, 5
[The spread of sexually transmitted diseases can be modelled by means of models that take into account the dynamic nature of social interactions.] [One aspect that is usually disregarded in the network approach is the dynamical nature of these interactions … epidemic.] [In this paper , we present a model of social dynamics that can be tailored to give similar accumulated degree distributions to those obtained in real surveys,] [but it also allows us to obtain some very general analytical results for the influence of network dynamics on the propagation of infectious diseases,] [using mean field techniques.]
Discourse Structure Tree of the Summary
Root
Background
N
S2-S3
S
Background
N
S2
LinkE
2
S3
N
Joint
N
3
EDU 4 & 5
N
Manner-Means
S
4
5
Nuclearity
Promotion Depth Score
S2: depth = 2
S3: depth = 4
Sentence Subtree Structure
Analysis Findings:
Finding 1: Sentences with complex structures are more prone to errors.
Finding 2: Errors are associated with the nuclearity and discourse features.
Section 4.1
2. Discourse Analysis on Summary
Discourse Parsed Source Document
Level 1
Level 2
Branched Structure
ASPL: 19.87
Chain-like Structure
ASPL: 17.02
Source Segmentation
Level 1: Segment 1: EDUs [1-648]
Segment 2: EDUs [649-688]
Level 2: Segments: [1-325]; [326-648]; [649-677]; [678-688]
Challenge: Current fact-checking models have a limit on the length of the context they can process.
- Prior work applies continuous chunking approach, which can disrupt the structure of the documents.
- We can leverage the parsed discourse tree to preserve the high-level integrity of segmented texts.
Original Document in GovReport dataset
Continuous Chunking
Discourse-inspired Segmentation
Section 4.2
3. Source Document Structure
Long Source Document
Model-generated Summary
(not factual)
RST Parser
0.91
0.46
0.43
Summary Sents
0.6
Discourse-inspired Chunks
Predicted summary-level label is incorrect
Section 5.1
Re-weighting Algorithm
1. Depth Score
2. Sentence Complexity
0.91
0.30
0.14
0.91
0.46
0.43
0.45
Section 5.2
4. Our Approach

Figure 1: Our proposed approach to faithfulness inconsistency detection utilizes findings from discourse analysis. We first conduct discourse analysis on parsed summary sentences (§4.1) and exploit the source document's discourse structure (§4.2). Motivated by the findings, our proposed approach is introduced in §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%
NoE	8%	23%	69%

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. Appendix C includes details of error types.

rizes our findings in §4.1 and §4.2.

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 $\sqrt{(28-24)} = 2$.[4]

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:

[4] 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
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 7 of Appendix C.

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).[5] 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. We compute both unnormalized and normalized versions (with the max tree depth) for the above three scores. As shown in Table 3, we find significant differences in the distributions of depth score. We normalize the 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 analyze 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, domain, 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 (Kim et al., 2024b). The intuition is that linear or chain-like graphs tend to have shorter average shortest path lengths (ASPL), reflecting 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.[6] In the meantime, longer summaries also carry evenly branched complex structures compared to short news summaries. While mainstream research segments long source texts into continuous chunks with limited window size, we argue that this disrupts the original structure of texts, leading to information loss.[7] We propose utilizing the tree structure and constructing the segments based on level traversals to preserve the high-level segmentation.

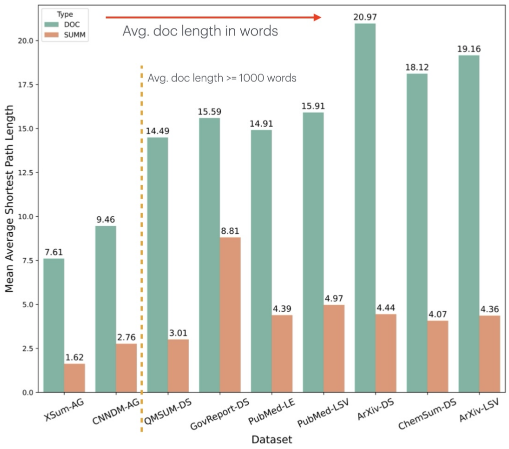


Figure 2: Average shortest path length per dataset for document and summary discourse trees. We sort the dataset by the average length of the document, finding that longer document-summary (DOC, SUMM) pairs would be more branched, and their summaries are also complicated. AG, DS, LSV, and LE refer to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

[5] Details of feature scores are in Appendix A.2.

[6] We exclude Multi-news and LEGALSUMM, as the former dataset's source text is composed of multiple news articles and the latter comes with split section structures, making the ASPL reporting less accurate.

[7] See Appendix D for examples.

However, as indicated in Section 4.1, EDUs with different discourse relations and structures can be 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 the associated discourse features 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).[8] 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 approach differs from prior work, which either applies a fixed window or selects 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 Experimental Details

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

Baselines One of our 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., 2024a), which sets the SOTA on DIVERSUMM, **MINICHECK FT5** (MiniCheck-FlanT5 checkpoints) (Tang et al., 2024) that is a best-performing 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 due to its design objectives of 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** (OpenAI et al., 2024) as the LLM fact-checker, using a prompt adopted from Tang et al. (2024) (see Table 8 in Appendix E). Lastly, we include **Llama-3.1-BeSpoke-**

[8] 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.

ID	Evaluation Model	AGGREFACT		DIVERSUMM						LSV		LONGEVAL	LEGALS
		XSM_{AG}	CND_{AG}	MNW	QMS	GOV	AXV	CSM	*Macro-*	PUB	AXV	PUB	
	evaluation metric	*AUC*				*AUC*			*AVG*	*Kendal's τ*		*Kendal's τ*	AUC
	avg src. len	360.54	518.85	669.20	1138.72	2008.16	4406.99	4612.40	–	3776.80	6236.40	3158.35	2873.87
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	60.19
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	61.33
3	**GPT4o**	75.36	70.47	51.11	70.22	86.81	67.78	61.53	67.49	54.7	51.8	51.2	67.71
4	**BeSpoke-MC-7B**	83.56	71.38	55.38	65.42	82.83	75.07	63.43	68.42	55.1	57.9	58.1	55.81
*Apply our approach with different **baselines**(↑ means improved the performance compared to the baseline with significance.)*													
5	**ALIGNSCORE**	75.66	69.50	46.74	56.48	87.02	77.46	61.03	65.75	54.9	53.9	36.9	73.52
6	+ re-weighting	75.67	69.20	45.33	53.95	87.29↑	81.15↑	60.55	65.65	53.0	54.3↑	34.8	76.57↑
7	+ LV1 SEGMENT	76.23↑	69.25†	45.86†	61.25↑	86.74†	79.47↑	64.15↑	67.49↑	51.9	52.8	43.6↑	59.43
8	STRUCTS-LV1	76.20↑	69.03	46.21†	60.06↑	86.04	82.78↑	64.47↑	67.91↑	50.4	53.9†	43.4↑	59.81
9	+ LV2 SEGMENT	74.27	70.30↑	46.03†	55.74	85.10	76.79	63.11↑	65.35	58.1↑	51.1	43.9↑	67.05
10	STRUCTS-LV2	74.28	69.85↑	45.33	51.86	85.65	80.00↑	63.59↑	65.29	55.3↑	54.1↑	43.7↑	64.00
11	**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	61.14
12	+ re-weighting	79.73	70.76†	56.79	60.36†	84.75↑	79.38↑	60.06↑	68.27↑	52.8	55.1↑	31.4↑	59.81
13	+ LV1 SEGMENT	77.84	73.48↑	44.80	61.10↑	87.50↑	85.22↑	63.59↑	68.44↑	57.5↑	51.4	33.0↑	68.95↑
14	STRUCTS-LV1	76.75	73.40↑	38.45	60.66†	88.05↑	86.32↑	63.11↑	67.31	56.2↑	53.8↑	30.7↑	72.57↑
15	+ LV2 SEGMENT	73.70	72.30↑	47.80	57.53	86.26↑	83.73↑	62.07↑	67.48	56.0↑	52.9↑	35.6↑	72.57↑
16	STRUCTS-LV2	71.31	72.30↑	41.27	59.02	87.16↑	84.78↑	61.75↑	66.80	53.4	54.2↑	33.0↑	73.71↑
17	**INFUSE**	68.48	72.52	54.14	39.64	84.41	68.13	57.82	60.83	59.4	55.9	36.9	63.43
18	+ re-weighting	67.30	72.37	53.44	40.54↑	84.68↑	74.31↑	59.82↑	62.56↑	58.3	56.3↑	34.6	66.29↑

Table 4: Results for all summarization tasks in AGGREFACT-FTSOTA (AGGREFACT), DIVERSUMM, LONGSCIVERIFY (LSV), LONGEVAL and LEGALSUMM (LegalS). 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 seven 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.

MiniCheck-7B (BeSpoke-MC-7B)[9], the SOTA fact-checking model on the LLM-AggreFact benchmark (Tang et al., 2024). Unless otherwise noted, we reran the baseline models on our datasets using the original authors' released code and checkpoints. Implementation details are provided in Appendix E.

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., 2024a), 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, DIVERSUMM and LEGALSUMM, following Laban et al. (2022); Zhang et al. (2024a), we adopt ROCAUC which measures classification performance with varied thresholds as our evaluation metric. On LONGSCIVERIFY and LONGEVAL, we report Kendall's Tau τ, following Bishop et al. (2024).

[9] https://huggingface.co/bespokelabs/Bespoke-MiniCheck-7B

7 Results

Overall Performance Table 4 presents our main results with detailed setups. Overall, our proposed approach (with different combinations of re-weighting and segmentation settings) achieves the best or second best across AGGREFACT, most of DIVERSUMM and LEGALSUMM (LEGALS). Compared to top-performed LLM-based models (rows 3,4), our approach outperforms in 7 out of 11 datasets, with significant improvements on GOV, AXV, CSM, and LEGALSUMM.[10] 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 5-6, 11-12, 17-18).* For long source documents, the re-weighting approach consistently improves or closely matches GOV, AXV, CSM splits in DIVERSUMM and the AXV split in LONGSCIVERIFY (LSV-AXV) and LEGALS performance. Noticeably, ALIGNSCORE with reweighting scored the best on LegalS. On the other hand, for both XSM and CND in AGGREFACT-FTSOTA, 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. 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 suggest 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, noticing both features are helpful, and the improvement in adding subtree height is greater.[11]

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 11 and 12 in Table 4.

RQ2. *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 (row 13) obtains the highest macro-average AUC on DIVERSUMM, a trend also observed with ALIGNSCORE. Specifically, comparing row 11 and row 13, the LV1 SEGMENT improved the model's performance on 7 of 8 long datasets from QMS to LEGALS (i.e. 78.66 -> 85.22 and 83.24 -> 87.50 on AXV and GOV). However, the effect of fine-grained segmentation can vary depending on the document's length and structure. For instance, ALIGNSCORE in row 9 with Lv2 segment obtained better performance than Lv1 on LSV-PUB but was worse 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 8 on AXV, CSM, and row 14 on AXV. 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 8 performs worse than row 5, likely due to the segmentation in row 7, making the model less accurate. Similarly, row 14 performs slightly better than row 11 on LSV-PUB, but row 13's improvement does not translate into better performance gains when combined with row 12. Differences in evaluation metrics (AUC vs. correlation) and dataset sizes may also have influenced these outcomes (i.e., row 14 does not improve much on LONGEVAL-PUB while rows 12 and 13 have larger gains).

8 Conclusion

In this work, we approach the factual inconsistency detection of long document summarization

[10] More discussions on strong baselines in Appendix F.1.

[11] We include a more complete table in Appendix F.2.

through the lens of discourse analysis. We find that discourse factors, with regard to sentence structure, are related to the factual consistency 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, achieving performance gains across multiple long-document summarization evaluation datasets, including scientific articles and legal documents.

Acknowledgment

This work is supported by the National Science Foundation under Grant No. 2040490, by Amazon, and by the University of Pittsburgh Center for Research Computing through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483. We want to thank the members of the Pitt PETAL group, Pitt NLP group, Pitt AI Fairness and Law group, and anonymous reviewers for their valuable comments in improving this work. We also thank Kevin D. Ashley and Morgan A. Gray for their help with data annotation.

Limitations

Our work contributes to the understanding of factual inconsistency errors in machine-generated summaries from the lens of discourse analysis. Here, we discuss several limitations.

Benefits of Discourse-driven Information Our current approach leaves discourse-relation information (i.e., the relation types such as Explanation, Elaboration, etc.) *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. We selected the current configuration that performed best on the validation splits of DIVERSUMM, aligning well with linguistic analysis principles. We plan to extend the modeling into a more complex version, such as applying a graph neural network to the tree structure and including discourse relations for future work.

While large models like GPT-4 and future architectures may improve long-context understanding, recent research shows that LLMs still face challenges with hallucination detection and effectively utilizing extended contexts (Liu et al., 2024a; Zhu et al., 2024; Luo et al., 2024). Our contribution, which links linguistic cues to hallucination detection, remains crucial, especially for summarization tasks. We acknowledge that future LLMs with expanded context windows may no longer require input pre-processing. However, we argue that discourse-based segmentation will *still offer critical benefits* (explicitly or implicitly by injecting the discourse analysis into the LLM through prompting or further finetuning). It will enhance the precision of factuality detection and evaluation by leveraging linguistic structures. Additionally, discourse information can provide interpretability to the model, which allows us to trace its evaluations to identifiable linguistic relations and features, which are still lacking in LLMs. In fact, our experimental results with BeSpoke-MC-7B, the SOTA fact-checking model, support the assumption that LLM alone still struggles with the factuality evaluation of long summaries.

Computation Cost Our approach's only additional computation cost is running the discourse parser on the source document and the target summary. The DMRST parser (Liu et al., 2021) can be run on both CPU and GPU, and the inference speed is fast (the full test set of DIVERSUMM can be processed in a few minutes). Once the discourse features are computed, the time spent by segmentation and reweighting algorithms remains static, introducing minimal overhead compared to the baselines.

Discourse-driven Analysis on Factual Errors 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).

Generalize across Text Domains We tried to cover most of the recent publicly available factuality evaluation datasets for long document summarization, including DiverSumm, LongSciVerify, and LongEval. While most existing datasets consist of annotations collected from scientific article summaries, we introduce a novel annotated dataset, LegalSumm, in the legal domain to evaluate the robustness of our proposed approaches. This dataset is curated with careful annotation procedures to ensure quality (see Appendix B). Our experimental results, as shown in the last column of Table 4, demonstrate that our proposed approaches not only enhance the performances of baseline models but also surpass those strong LLM-based models by a large margin.

Dependence on Discourse Parser Performance Our experiments' validity and subsequent findings rely on the parsed discourse trees generated by a Rhetorical Structure Theory (RST) parser (Liu et al., 2021), following prior work (Adams et al., 2023a; Pu et al., 2023; Kim et al., 2024b). It is important to note that parsed results may 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 acknowledge that RST parsers are gradually evolving and posit that better RST parsing results can further boost the model's performance. However, major obstacles to their broader adoption are the lack of publicly available models and user-friendly user guidance. Researchers recently incorporated LLMs in discourse parsing and obtained better benchmarking performance in RST (Maekawa et al., 2024). Unfortunately, no available inference code exists to parse documents beyond pre-compiled benchmark datasets. We look forward to utilizing more robust parsers in future work.

On long source documents, we notice that the parser failed on the MNW split of the DiverSumm, given their input is a concatenation of multiple individual news articles. We opt for first splitting the original document into articles and then successfully parsing them individually. Regarding paragraph-level discourse parsing, we are concerned that it may disrupt the discourse continuity at the document level (i.e., where the beginning of one paragraph is connected to the previous paragraph). Therefore, we leave this exploration for future studies. However, this approach might be viable and beneficial for summarizing extremely long documents, such as books, where the explicit division into chapters and sections could enhance the process.

Applications of Document Structures to Other Tasks Document structures can and have been utilized in different tasks, including coherence analysis (Liu et al., 2024b), machine translation evaluation (Joty et al., 2017), sentiment analysis (Kraus and Feuerriegel, 2019), machine-generated text detection (Kim et al., 2024b), etc. While applying document structure and discourse analysis to hallucination detection is still an emerging area of research, we are keen to explore it further. We are also interested in extending this approach to other input sources, such as dialogue, by investigating the corresponding discourse structures unique to conversational data.

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.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. Internlm2 technical report.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

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.

Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. 2002. Rst discourse treebank. Technical report, Linguistic Data Consortium, University of Pennsylvania.

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.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mohamed Elaraby, Huihui Xu, Morgan Gray, Kevin Ashley, and Diane Litman. 2024. Adding argumentation into human evaluation of long document abstractive summarization: A case study on legal opinions. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 28–35, Torino, Italia. ELRA and ICCL.

Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Towards argument-aware abstractive summarization of long legal opinions with summary reranking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7601–7612, Toronto, Canada. 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.

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.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2017. Discourse structure in machine translation evaluation. *Computational Linguistics*, 43(4):683–722.

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. *First Conference on Language Modeling*.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024b. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

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.

Mathias Kraus and Stefan Feuerriegel. 2019. Sentiment analysis based on rhetorical structure theory:learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:65–79.

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.

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

Yinhong Liu, Yixuan Su, Ehsan Shareghi, and Nigel Collier. 2024b. Unlocking structure measuring: Introducing PDD, an automatic metric for positional discourse coherence. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 92–100, Mexico City, Mexico. Association for Computational Linguistics.

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.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2803–2815, St. Julian's, Malta. 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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

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.

Huihui Xu, Jaromir Savelka, and Kevin D. Ashley. 2021. Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. In *International Conference on Legal Knowledge and Information Systems*.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Miami, Florida, USA. Association for Computational Linguistics.

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. 2024a. 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. 2024b. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59670–59684. PMLR.

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.

Zhiying Zhu, Yiming Yang, and Zhiqing Sun. 2024. Halueval-wild: Evaluating hallucinations of language models in the wild.

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).[12] 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 Features

Following prior work (Louis et al., 2010), we analyze the nucleus-satellite penalty score (Ono penalty) (Ono et al., 1994), the maximum depth (Depth score) (Marcu, 1998), and the promotion-based score (Marcu, 1998) for sentence level. The penalty/score for a sentence is computed as the maximum of the penalties/scores of its constituent EDUs. For the normalized version, instead of following Louis et al. (2010), who normalized them by the number of words in the document, we opt to divide the scores by the maximum depth of the discourse tree, which similarly alleviates the scores' dependencies on document length. Below, we provide one example demonstrating the computation of each score (borrowed from Louis et al. (2010)) and will release our code for reproduction purposes.

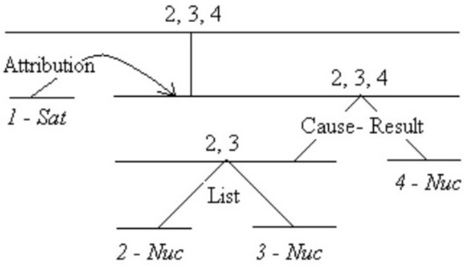


Figure 3: RST for the example sentence, and the salient units (promotion set) of each text span are shown above the horizontal line, which represents the span.The example is taken from Louis et al. (2010).

A.2.1 Example

Here, we re-utilize the example from Louis et al. (2010), which is part of the RSTDT (Carlson et al., 2002) in Figure 3, which contains four EDUs.

1. [Mr. Watkins said] 2. [volume on Interprovincial's system is down about 2% since January] 3. [and is expected to fall further,] 4. [making expansion unnecessary until perhaps the mid-1990s.]

Nucleaus-Satellite Penalty (Ono Penalty) (Ono et al., 1994): The spans of individual EDUs are represented at the leaves of the tree. At the root of the tree, the span covers the entire text. The path from EDU 1 to the root contains one satellite node. It is, therefore, assigned a penalty of 1. Paths to the root from all other EDUs involve only nucleus nodes; subsequently, these EDUs do not incur any penalty. Thus, the Ono Penalty scores for EDU 1 to 4 are [1, 0, 0, 0].

Maximum Depth Score Below we cite the original texts from (Louis et al., 2010).

> Marcu (1998) proposed the method to utilize the nucleus-satellite distinction, rewarding nucleus status instead of penalizing the satellite. He introduced the notion of *promotion set*, consisting of salient/important units of a text span. The nucleus is denoted as the more salient unit in the full span of a mononuclear relation (i.e., in Elaboration, the satellite unit is to elaborate on the key information of the nucleus. Thus, the latter is more salient). In a multinuclear

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

relation, all the nuclei are salient units of the larger span.

For example, in Figure 3, EDUs 2 and 3 participate in a multinuclear (List) relation. As a result, both EDUs 2 and 3 appear in the promotion set of their combined span (2-3). The salient units (promotion set) of each text span are shown above the horizontal line which represents the span. At the leaves, salient units are the EDUs themselves.

For the purpose of identifying important content, units in the promotion sets of nodes close to the root are hypothesized to be more important than those at lower levels. The highest promotion of an EDU occurs at the node closest to the root, which contains that EDU in its promotion set. The depth of the tree from the highest promotion is assigned as the score for that EDU. Hence, the closer to the root an EDU is promoted, the better its score. Since EDUs 2, 3 and 4 are promoted all the way up to the root of the tree, the score assigned to them is equal to 4, the total depth of the tree. EDU 1 receives a depth score of 3.

Thus, the final maximum depth score based on the promotion set for EDUs 1-4 are [3, 4, 4, 4].

Promotion Score In the same example, while EDUs 2, 3, and 4 all have a depth score of 4, EDUs 2 and 3 are promoted to the root from a greater depth than EDU 4. To account for the difference, Marcu (1998) further introduced the promotion score, which is a measure of the number of levels over which an EDU is promoted. For instance, EDU 2 is promoted by three levels, while EDU 4 is promoted by two levels. Thus, EDUs 2 and 3 receive a promotion score of 3, while the score of EDU 4 is only 2. EDU 1, given that it is never promoted received scores of 0.

Discourse Tree Computation In Section 4 Table 2, we compute the tree depth as follows. We use a string-matching system to construct a dictionary that aligns annotated sentences with EDU segments. For instance, in Figure 3, the sentence is mapped to EDUs 1-4. We then compute the maximum depth of the discourse tree from the root node to the lowest leaf node, which would be 3 in this case. However, there may be cases where sentences are segmented into EDUs that are not gathered into a single node in the parsed discourse tree. In such instances, we employ the methods described in Section 4 to approximate the depth.

B LegalSumm Dataset

We utilized a subset of the **CanLII Dataset** (Xu et al., 2021), which consisted of 1,049 legal opinion documents with expert-written summaries.[13]. We followed the setting from Elaraby et al. (2024), where we consider the output of three different abstractive models in our annotation process: (1) **Finetuned LED-base** (Elaraby and Litman, 2022) which finetuned the pre-trained longformer-encoder-decoder (Beltagy et al., 2020) (LED) on the CanLII cases without additional information about the argument structure of the document (2) **arg-LED-base**, which utilizes the LED model but includes the information about the argument units (Issues, Reasons, and Conclusions) in its training phase, and (3) **arg-aug-LED-base**, a model introduced in Elaraby et al. (2023) that can select a summary from multiple augmented versions of generated summaries based on its overlap with the input case's predicted argument roles.

Annotation Details We conducted evaluations with two voluntary legal experts from the research group, all of whom hold a J.D. degree and possess at least four years of experience in providing professional legal services. For each summary, the annotators are asked to select from four choices justifying the factual consistency of the model-generated summary with the reference summary and source article. They are also encouraged to provide free-text rationales justifying their selections.

To guarantee the quality of annotation, we conducted multiple sessions with annotators to refine the guidelines and continuously monitor the agreements. Ultimately, the first author and the two annotators held in-person sessions to resolve label inconsistencies. The labels remained unresolved in two cases as the annotators identified differing yet reasonable interpretations of the instructions. We thus retain the average scores as is. To distinguish summaries with severe or moderate factual inconsistencies from those without, we computed the average of the two annotators' ratings and rounded

[13] Data obtained through an agreement with CanLII https://www.canlii.org/en/

based on a threshold of 2. The annotation guideline is included in Figure 4.

C Discourse Analysis on Fine-grained Error Types

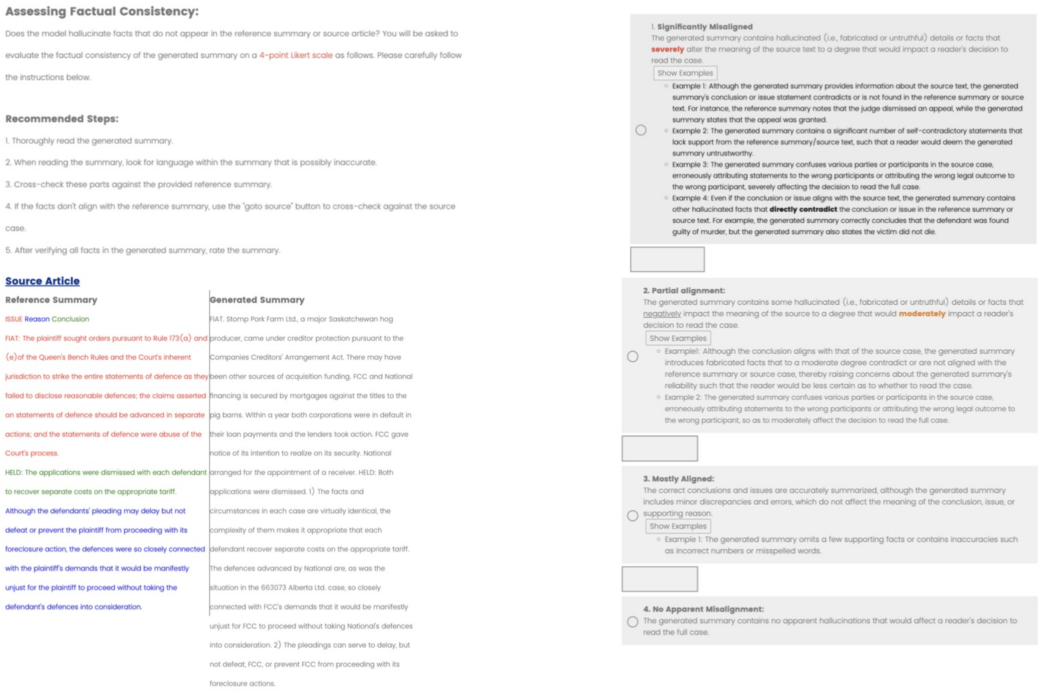
Error Types Relation Error (PreE) is when the predicate in a summary sentence is inconsistent with respect to the document. Entity Error (EntE) is when the primary arguments of the predicate are incorrect. Circumstance Error (CircE) is when the predicate's circumstantial information (i.e., name or time) 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 7, we demonstrate the breakdowns of fine-grained error types and report the t-test results on different discourse features.

D 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 5a, 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 6a, the conclusion section is separated into two chunks by the default chunking approach, while our method maintains them in a single chunk.



Assessing Factual Consistency:

Does the model hallucinate facts that do not appear in the reference summary or source article? You will be asked to evaluate the factual consistency of the generated summary on a 4-point Likert scale as follows. Please carefully follow the instructions below.

Recommended Steps:

1. Thoroughly read the generated summary.
2. When reading the summary, look for language within the summary that is possibly inaccurate.
3. Cross-check these parts against the provided reference summary.
4. If the facts don't align with the reference summary, use the "goto source" button to cross-check against the source case.
5. After verifying all facts in the generated summary, rate the summary.

Source Article

Reference Summary

ISSUE Reason Conclusion

FIAT: The plaintiff sought orders pursuant to Rule 173(a) and (e)of the Queen's Bench Rules and the Court's inherent jurisdiction to strike the entire statements of defence as they failed to disclose reasonable defences; the claims asserted on statements of defence should be advanced in separate actions; and the statements of defence were abuse of the Court's process.

HELD: The applications were dismissed with each defendant to recover separate costs on the appropriate tariff.

Although the defendants' pleading may delay but not defeat or prevent the plaintiff from proceeding with its foreclosure action, the defences were so closely connected with the plaintiff's demands that it would be manifestly unjust for the plaintiff to proceed without taking the defendant's defences into consideration.

Generated Summary

FIAT. Stomp Pork Farm Ltd., a major Saskatchewan hog producer, came under creditor protection pursuant to the Companies Creditors' Arrangement Act. There may have been other sources of acquisition funding. FCC and National financing is secured by mortgages against the titles to the pig barns. Within a year both corporations were in default in their loan payments and the lenders took action. FCC gave notice of its intention to realize on its security. National arranged for the appointment of a receiver. HELD: Both applications were dismissed. 1) The facts and circumstances in each case are virtually identical, the complexity of them makes it appropriate that each defendant recover separate costs on the appropriate tariff. The defences advanced by National are, as was the situation in the 663073 Alberta Ltd. case, so closely connected with FCC's demands that it would be manifestly unjust for FCC to proceed without taking National's defences into consideration. 2) The pleadings can serve to delay, but not defeat, FCC, or prevent FCC from proceeding with its foreclosure actions.

○ 1. **Significantly Misaligned**
The generated summary contains hallucinated (i.e., fabricated or untruthful) details or facts that **severely** alter the meaning of the source text to a degree that would impact a reader's decision to read the case.
Show Examples
- Example 1: Although the generated summary provides information about the source text, the generated summary's conclusion or issue statement contradicts or is not found in the reference summary or source text. For instance, the reference summary notes that the judge dismissed an appeal, while the generated summary states that the appeal was granted.
- Example 2: The generated summary contains a significant number of self-contradictory statements that lack support from the reference summary/source text, such that a reader would deem the generated summary untrustworthy.
- Example 3: The generated summary confuses various parties or participants in the source case, erroneously attributing statements to the wrong participants or attributing the wrong legal outcome to the wrong participant, severely affecting the decision to read the full case.
- Example 4: Even if the conclusion or issue aligns with the source text, the generated summary contains other hallucinated facts that **directly contradict** the conclusion or issue in the reference summary or source text. For example, the generated summary correctly concludes that the defendant was found guilty of murder, but the generated summary also states the victim did not die.

○ **2. Partial alignment:**
The generated summary contains some hallucinated (i.e., fabricated or untruthful) details or facts that negatively impact the meaning of the source to a degree that would **moderately** impact a reader's decision to read the case.
Show Examples
- Example1: Although the conclusion aligns with that of the source case, the generated summary introduces fabricated facts that to a moderate degree contradict or are not aligned with the reference summary or source case, thereby raising concerns about the generated summary's reliability such that the reader would be less certain as to whether to read the case.
- Example 2: The generated summary confuses various parties or participants in the source case, erroneously attributing statements to the wrong participants or attributing the wrong legal outcome to the wrong participant, so as to moderately affect the decision to read the full case.

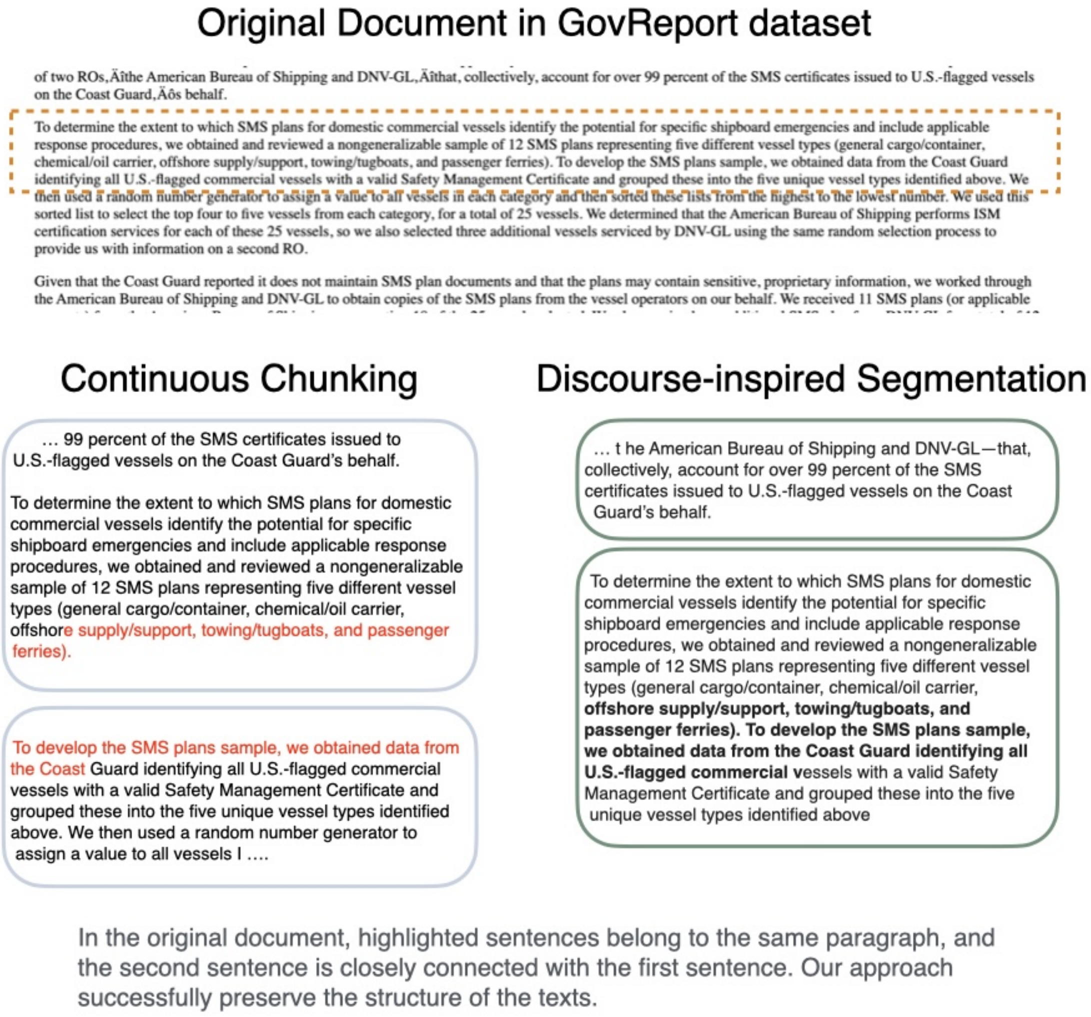
○ **3. Mostly Aligned:**
The correct conclusions and issues are accurately summarized, although the generated summary includes minor discrepancies and errors, which do not affect the meaning of the conclusion, issue, or supporting reason.
Show Examples
- Example 1: The generated summary omits a few supporting facts or contains inaccuracies such as incorrect numbers or misspelled words.

○ **4. No Apparent Misalignment:**
The generated summary contains no apparent hallucinations that would affect a reader's decision to read the full case.

Figure 4: The annotation interface for LegalSumm. The left panel displays the instructions and the content to be annotated. Annotators are then prompted to select one of four options, as shown in the right panel.

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 7: 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.



Original Document in GovReport dataset

of two ROs,Äîthe American Bureau of Shipping and DNV-GL,Äîthat, collectively, account for over 99 percent of the SMS certificates issued to U.S.-flagged vessels on the Coast Guard,Äôs behalf.

To determine the extent to which SMS plans for domestic commercial vessels identify the potential for specific shipboard emergencies and include applicable response procedures, we obtained and reviewed a nongeneralizable sample of 12 SMS plans representing five different vessel types (general cargo/container, chemical/oil carrier, offshore supply/support, towing/tugboats, and passenger ferries). To develop the SMS plans sample, we obtained data from the Coast Guard identifying all U.S.-flagged commercial vessels with a valid Safety Management Certificate and grouped these into the five unique vessel types identified above. We then used a random number generator to assign a value to all vessels in each category and then sorted these lists from the highest to the lowest number. We used this sorted list to select the top four to five vessels from each category, for a total of 25 vessels. We determined that the American Bureau of Shipping performs ISM certification services for each of these 25 vessels, so we also selected three additional vessels serviced by DNV-GL using the same random selection process to provide us with information on a second RO.

Given that the Coast Guard reported it does not maintain SMS plan documents and that the plans may contain sensitive, proprietary information, we worked through the American Bureau of Shipping and DNV-GL to obtain copies of the SMS plans from the vessel operators on our behalf. We received 11 SMS plans (or applicable

Continuous Chunking

… 99 percent of the SMS certificates issued to U.S.-flagged vessels on the Coast Guard's behalf.

To determine the extent to which SMS plans for domestic commercial vessels identify the potential for specific shipboard emergencies and include applicable response procedures, we obtained and reviewed a nongeneralizable sample of 12 SMS plans representing five different vessel types (general cargo/container, chemical/oil carrier, offshore supply/support, towing/tugboats, and passenger ferries).

To develop the SMS plans sample, we obtained data from the Coast Guard identifying all U.S.-flagged commercial vessels with a valid Safety Management Certificate and grouped these into the five unique vessel types identified above. We then used a random number generator to assign a value to all vessels I ….

Discourse-inspired Segmentation

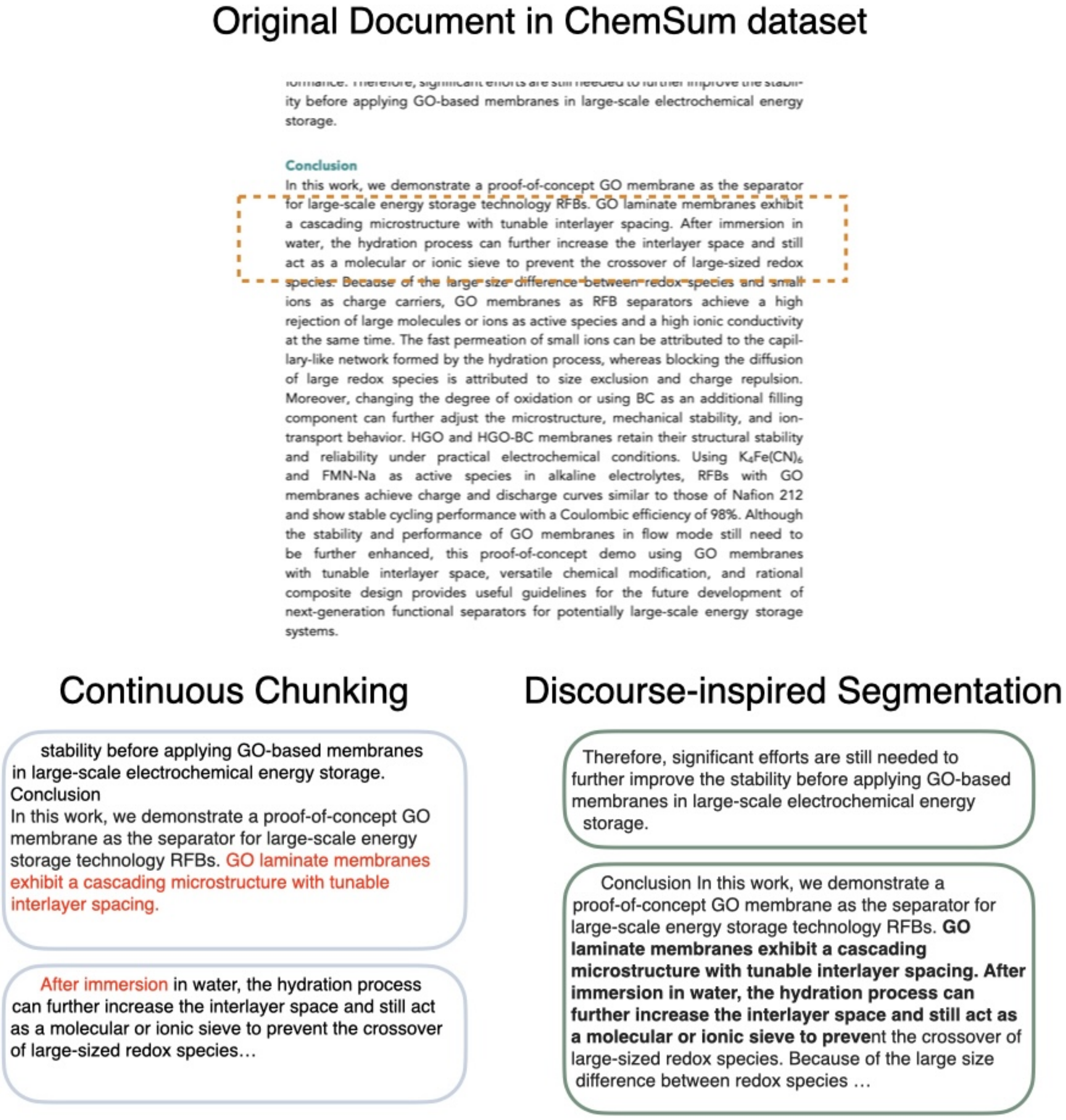
… t he American Bureau of Shipping and DNV-GL—that, collectively, account for over 99 percent of the SMS certificates issued to U.S.-flagged vessels on the Coast Guard's behalf.

To determine the extent to which SMS plans for domestic commercial vessels identify the potential for specific shipboard emergencies and include applicable response procedures, we obtained and reviewed a nongeneralizable sample of 12 SMS plans representing five different vessel types (general cargo/container, chemical/oil carrier, **offshore supply/support, towing/tugboats, and passenger ferries). To develop the SMS plans sample, we obtained data from the Coast Guard identifying all U.S.-flagged commercial v**essels with a valid Safety Management Certificate and grouped these into the five unique vessel types identified above

In the original document, highlighted sentences belong to the same paragraph, and the second sentence is closely connected with the first sentence. Our approach successfully preserve the structure of the texts.

(a) Example from GovReport of DIVERSUMM.

Figure 5: 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.



Original Document in ChemSum dataset

formance. Therefore, significant efforts are still needed to further improve the stability before applying GO-based membranes in large-scale electrochemical energy storage.

Conclusion

In this work, we demonstrate a proof-of-concept GO membrane as the separator for large-scale energy storage technology RFBs. GO laminate membranes exhibit a cascading microstructure with tunable interlayer spacing. After immersion in water, the hydration process can further increase the interlayer space and still act as a molecular or ionic sieve to prevent the crossover of large-sized redox species. Because of the large size difference between redox species and small ions as charge carriers, GO membranes as RFB separators achieve a high rejection of large molecules or ions as active species and a high ionic conductivity at the same time. The fast permeation of small ions can be attributed to the capillary-like network formed by the hydration process, whereas blocking the diffusion of large redox species is attributed to size exclusion and charge repulsion. Moreover, changing the degree of oxidation or using BC as an additional filling component can further adjust the microstructure, mechanical stability, and ion-transport behavior. HGO and HGO-BC membranes retain their structural stability and reliability under practical electrochemical conditions. Using $K_4Fe(CN)_6$ and FMN-Na as active species in alkaline electrolytes, RFBs with GO membranes achieve charge and discharge curves similar to those of Nafion 212 and show stable cycling performance with a Coulombic efficiency of 98%. Although the stability and performance of GO membranes in flow mode still need to be further enhanced, this proof-of-concept demo using GO membranes with tunable interlayer space, versatile chemical modification, and rational composite design provides useful guidelines for the future development of next-generation functional separators for potentially large-scale energy storage systems.

Continuous Chunking

stability before applying GO-based membranes in large-scale electrochemical energy storage.
Conclusion
In this work, we demonstrate a proof-of-concept GO membrane as the separator for large-scale energy storage technology RFBs. GO laminate membranes exhibit a cascading microstructure with tunable interlayer spacing.

After immersion in water, the hydration process can further increase the interlayer space and still act as a molecular or ionic sieve to prevent the crossover of large-sized redox species…

Discourse-inspired Segmentation

Therefore, significant efforts are still needed to further improve the stability before applying GO-based membranes in large-scale electrochemical energy storage.

Conclusion In this work, we demonstrate a proof-of-concept GO membrane as the separator for large-scale energy storage technology RFBs. **GO laminate membranes exhibit a cascading microstructure with tunable interlayer spacing. After immersion in water, the hydration process can further increase the interlayer space and still act as a molecular or ionic sieve to preven**t the crossover of large-sized redox species. Because of the large size difference between redox species …

(a) Example from ArXiv of DIVERSUMM.

Figure 6: 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.

E Implementation Details

E.1 GPT4o Prompts

We include our prompt for zero-shot factual consistency evaluation in Table 8.

E.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 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 sentences. 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. (2024a) 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 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. 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" answers as the summary-level rating.

BeSpoke-MC-7B We harnessed the SOTA Llama-3.1-Bespoke-MiniCheck-7B (BeSpoke-MC-7B) released by Bespoke Labs. The model is fine-tuned from "internlm/internlm2_5-7b-chat" (Cai et al., 2024) on the combination of 35K data points following the approach in MiniCheck (Tang et al., 2024). We use the suggested code repo from https://huggingface.co/bespokelabs/Bespoke-MiniCheck-7B. To calculate the AUC score, we employed the raw probabilities returned by the code to determine sentence-level ratings, and we calculated the summary-level score as the unweighted average across all sentences.

E.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[14] (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. Additionally, we loaded the Bespoke-MC-7B model

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

```
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 8: Zero-shot factual consistency evaluation prompt for GPT4o.

on a single NVIDIA L40S GPU with 48 GB of VRAM, provided by the Pitt CRC computing cluster.

F Experimental Results

F.1 Discussion on Performance Compared to Strong Baselines

Our primary analysis focuses on discussing how the proposed approach can improve different baselines (we utilized three backbone baselines: rows 5, 11, and 17 with their improved versions) in Table 4. We observe several baselines obtained the best performance on certain tasks and provide a more careful justification below:

While the improvements may appear marginal in some baseline models, they are statistically significant and consistent across multiple datasets. The capabilities of baseline models and the characteristics of testbeds can also affect performance. For instance, as noted in Section 7, dialogue-based inputs in QMS limit the effectiveness of discourse parsing (RQ2), while short summaries like XSUM minimize the impact of reweighting (RQ1). On longer datasets like AXV and CSM, gains are more substantial, with improvements of up to 7 points (row 14 vs. row 11 in AXV). This is comparable to, or even more significant than, prior work (Zhang et al., 2024a), and it is common to observe varying levels of performance gains across different tasks (Tang et al., 2023, 2024).

LongDocFactScore (LDFS) introduced the LongSciVerify (LSV) dataset (PUB and AXV), using a different annotation method by subsampling three sentences with human annotations for factuality. We conjecture this may lead to less accurate summary-level labels, favoring their metric, which utilizes the top-k sentence-level scores. Meanwhile, LDFS underperformed compared to most other baselines on AggreFact, QMS, AXV (from DIVERSUMM), and LongEval-PUB. In contrast, our approach outperformed LongDocFactScore on most other benchmarks (e.g., 86.32 vs. 65.36 on AXV), suggesting our approach is more robust and capable of handling different long document summarization datasets. While each baseline may excel in specific tasks, a more robust benchmarking dataset could better ensure fair comparisons for future research.

GPT-4o GPT4o is utilized as a comparison between the SOTA LLMs (GPT4o models have unknown sizes but could be greater than known open-sourced LLMs with up to 405B) and our lightweight model (770M), which in the usual case, the LLMs can outperform baselines by noticeable margins (Tang et al., 2024)). In Table 4, regarding the long document summarization datasets (from GOV in DiverSumm to LegalSumm), our models (rows 12, 13) outperformed GPT4o in 5 out of 6 test sets (the only exception is LongEval PUB). This confirmed that the discourse-inspired approaches are beneficial.

BeSpoke-MC-7B is claimed to be the best fact-checking model publicly available on the LLM-AggreFact benchmark, which outperformed many other LLMs with bigger sizes. Compared to our proposed models, it performed better on QMS, XSM_{AG}, LSV-AXV, and had the best performance on LongEval-PUB (similar pattern to GPT-4o). However, on other benchmarks, our discourse approaches still demonstrate their benefits (i.e., on LEGALSUMM, AlignScore + reweighting obtained 76.57 while BeSpoke-MC-7B only scored 55.81).

F.2 Ablation Study

Table 9 presents the ablation results of different discourse features on our baselines. We cover the long document summarization tasks starting from QMS in Table 4.

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 9: Ablation results on long document datasets from DiverSumm, LongSciVerify and LongEval.