Structure-Aware Chunking for Abstractive Summarization of Long Legal Documents

Himadri Sonowal[1] [*] **Saisab Sadhu**[1]
[1]Department of Data Science and Engineering,
Indian Institute of Science Education and Research Bhopal, India
adrisonowal@gmail.com, sadhusaisab@gmail.com

Abstract

The efficacy of state-of-the-art abstractive summarization models is severely constrained by the extreme document lengths of legal judgments, which consistently surpass their fixed input capacities. The prevailing method, naive sequential chunking, is a discourse-agnostic process that induces context fragmentation and degrades summary coherence. This paper introduces **Structure-Aware Chunking (SAC)**, a rhetorically-informed pre-processing pipeline that leverages the intrinsic logical structure of legal documents. We partition judgments into their constituent rhetorical strata—Facts, Arguments & Analysis, and Conclusion—prior to the summarization pass. We present and evaluate two SAC instantiations: a computationally efficient heuristic-based segmenter and a semantically robust LLM-driven approach. Empirical validation on the JUST-NLP 2025 L-SUMM shared task dataset reveals a nuanced trade-off: while our methods improve local, n-gram based metrics (ROUGE-2), they struggle to maintain global coherence (ROUGE-L). We identify this "coherence gap" as a critical challenge in chunk-based summarization and show that advanced LLM-based segmentation begins to bridge it. To facilitate reproducibility, we release our code and pre-processing scripts.[1]

1 Introduction

Automated summarization of legal documents is a critical task for managing information overload in digital archives. Abstractive summarization, which aims to generate fluent and concise synopses of complex documents, is a promising avenue for improving the efficiency of legal systems and enhancing access to justice. However, a fundamental granularity mismatch severely limits its application to the legal domain: the length of typical court judgments, which often exceed 10,000 tokens (Shukla et al., 2022), is frequently orders of magnitude larger than the token capacity of state-of-the-art transformer models (Lewis et al., 2020; Zhang et al., 2020). While recent architectural innovations have enabled the processing of much longer documents (Bashir et al., 2025; Chhibbar and Kalita, 2024), these approaches still face challenges in maintaining coherence across ultra-long legal texts (Moro et al., 2023).

Because of this discrepancy, ultra-long documents must be pre-processed into chunks that can be ingested by models. The prevailing technique, which we refer to as Naive Sequential Chunking (NSC), divides the document into fixed-size, non-overlapping blocks implementing a brute-force segmentation. Despite its simplicity, this method is conceptually flawed because it disregards the document's discourse structure. Legal judgments are highly structured with a canonical rhetorical progression: Facts, Arguments & Analysis, and Conclusion. These logical units are randomly split by NSC, which breaks cohesive sequences of reasoning and separates premises from their conclusions. As a result, the context becomes fragmented, leading to anaphora resolution failures (Steinberger et al., 2007) and a disjointed final summary. This fragmentation is a key manifestation of the "coherence gap" we investigate.

To address this, we propose **Structure-Aware Chunking (SAC)**, a pipeline that aligns the chunking process with the document's rhetorical schema. We implement and evaluate two methods for this segmentation: a lightweight heuristic-based approach (SAC-H) and a semantically robust, zero-shot LLM-based approach (SAC-LLM). Our contribution is not merely the proposal of a new method, but a systematic investigation that uncovers a critical and counter-intuitive trade-off between local fluency (e.g., ROUGE-2) and global coherence (e.g., ROUGE-L) in chunk-based summarization of long, structured documents.

[*] Corresponding author.
[1] https://github.com/sonowalh/sac-legal-summ

Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 171–178
December 24, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Our research is situated at the intersection of long-document summarization, legal NLP, and evaluation methodologies.

2.1 Long-Document and Chunking Strategies

The challenge of processing documents that exceed model input capacity has motivated diverse strategies. Architectural innovations, such as the efficient attention mechanisms in Longformer (Beltagy et al., 2020), represent one major line of inquiry. Another involves hierarchical models, which first encode smaller text units before aggregating them to form a document-level understanding (Sun et al., 2024; Wang et al., 2024). A third paradigm is the hybrid extractive-abstractive approach, where an extractive stage creates a compressed intermediate document for abstractive synthesis (Divya et al., 2024; Datta et al., 2023). Our approach is orthogonal to these, as SAC is a pre-processing strategy that can be integrated with any of these model types by operating at a discourse level.

The necessity of pre-processing has led to a focus on chunking strategies (Kumar et al., 2024). While fixed-size, discourse-agnostic chunking remains common (Pinecone, 2025), more advanced methods have explored sentence-aware segmentation (Miculicich and Han, 2023). Miculicich and Han (2023) provide empirical support for our premise, demonstrating that incorporating text segmentation improves extractive summarization by reducing lead bias. Furthermore, the exploration of sliding windows with overlap to improve local context continuity (Koay et al., 2021) directly informs our **SAC-H+** variant, which adapts this concept to operate within rhetorical boundaries.

2.2 Rhetorical Structure in Legal NLP

In the legal domain, segmentation can be informed by the text's well-established rhetorical structure. Recent advances in legal NLP have established robust frameworks for rhetorical role classification. Nigam et al. (2025) introduced LegalSeg, the largest annotated dataset of its kind, demonstrating that models incorporating structural relationships outperform sentence-level approaches. Earlier work (Hachey and Grover, 2004; Bhattacharya et al., 2019) established the foundations for such classification, while transformer-based approaches (Marino et al., 2023; Joshi et al., 2024) have recently achieved state-of-the-art performance. These developments validate our premise that leveraging rhetorical structure is crucial. However, prior work has focused primarily on extractive summarization and classification. Our work bridges this gap by being the first, to our knowledge, to leverage rhetorical structure as a pre-processing strategy specifically for *abstractive synthesis* of ultra-long legal texts.

2.3 Evaluation of Summarization

The evaluation of summarization has moved beyond simple n-gram overlap metrics like ROUGE (Lin, 2004). While ROUGE remains a dominant evaluation metric, its focus on lexical overlap has known limitations for assessing semantic quality and factual consistency (Kryściński et al., 2020). Modern standards emphasize semantic similarity via contextual embeddings, with BERTScore (Zhang et al., 2019) becoming a de facto standard. For high-stakes domains like law, factual consistency metrics (Elaraby et al., 2024; Luo et al., 2024) and discourse coherence models (Zhao et al., 2023; Lin et al., 2011) are also gaining prominence. Informed by this, our evaluation strategically employs ROUGE-L as a proxy for global coherence, contrasting it with ROUGE-2 for local fluency, to investigate the trade-offs inherent in chunk-based summarization.

3 Methodology

Our methodology is designed as a multi-stage pipeline that transforms a raw, ultra-long document into a coherent summary, as depicted in Figure 1. The core innovation lies in the first two stages: Rhetorical Segmentation and Proportional Budget Allocation. We first describe our experimental setup and baseline before detailing the SAC pipeline.

3.1 Implementation Details

For all experiments, we use Legal-Pegasus (Sharma et al., 2023), a Pegasus model pre-trained on a large corpus of legal text. We utilize the `nsi319/legal-pegasus` checkpoint, which imposes a maximum input sequence length of $n_{max} = 1024$ tokens.

All experiments were conducted on a single NVIDIA H100 GPU. The Legal-Pegasus model was used with a beam size of 4, a length penalty of 2.0, and a repetition penalty of 1.2. For our

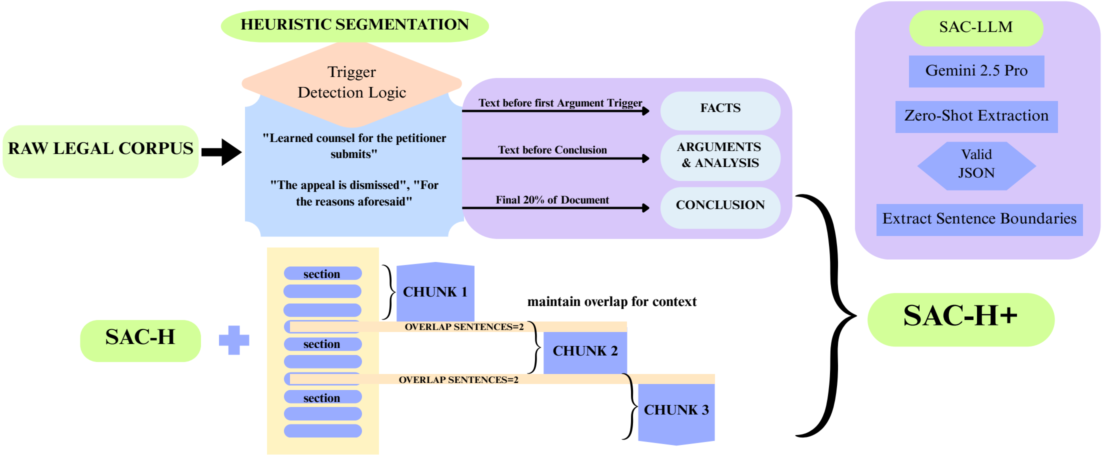

Figure 1: The Structure-Aware Chunking (SAC) Pipeline. A long document is first segmented into rhetorical sections, a summary budget is allocated to each, and then each section is chunked and summarized before final concatenation.

SAC-LLM method, we utilized Gemini 2.5 Pro[2] accessed via the OpenRouter API[3].

3.2 Baseline: Naive Sequential Chunking (NSC)

Our baseline, NSC, partitions a document D into a sequence of $k = \lceil |D|/n_{max} \rceil$ non-overlapping chunks $\{C_1, \ldots, C_k\}$. The final summary S is a concatenation of the sub-summaries $S_i = \text{Summarize}(C_i)$, where the target length of each S_i is uniformly set to L_{target}/k.

3.3 Proposed Method: Structure-Aware Chunking (SAC)

The SAC pipeline consists of two main stages, detailed below.

3.3.1 Stage 1: Rhetorical Segmentation

This stage partitions the document D into three canonical legal sections: Facts (D_{facts}), Arguments & Analysis ($D_{arg\&an}$), and Conclusion (D_{conc}). We implement and compare two methods for this stage.

SAC-Heuristic (SAC-H). This method employs a computationally efficient, top-down cascade of heuristic rules based on high-precision lexical triggers. The process is as follows:

1. **Conclusion Identification:** The algorithm first anchors the segmentation by identifying the final ruling. Based on an empirical analysis of 50 randomly sampled documents from the training set, which showed that 95% of conclusions appear in the final 20% of the document, our search is constrained to this region. It identifies the first instance of conclusive phrases (e.g., "The appeal is dismissed").

2. **Arguments & Analysis Identification:** It then searches the text preceding the identified conclusion for phrases signaling the onset of legal argumentation (e.g., "Learned counsel for the petitioner submits").

3. **Section Delineation:** The text before the first argument trigger constitutes D_{facts}, the text between it and the conclusion trigger forms $D_{arg\&an}$, and the final part is D_{conc}. A comprehensive list of the trigger phrases is provided in Appendix B

SAC-Heuristic+ (SAC-H+). As an enhancement to SAC-H, we introduce SAC-H+, which addresses potential context fragmentation within long rhetorical sections. While SAC-H correctly delineates the major rhetorical strata, a very long "Arguments & Analysis" section might still be split into multiple chunks. To mitigate the hard boundary effects of this intra-section chunking, SAC-H+ incorporates a sentence-aware sliding window with a 2-sentence overlap. When a section D_{sec} is chunked, each subsequent chunk C_i begins with the final two sentences of the preceding chunk C_{i-1}. This provides the model with local contextual continuity, aiming to improve the flow between the sub-summaries generated from a single rhetorical block.

[2] https://deepmind.google/models/gemini/pro/
[3] https://openrouter.ai/

SAC-LLM. Our zero-shot LLM approach leverages the large context window of Gemini 2.5 Pro to maintain global document awareness during rhetorical boundary identification. The prompt is structured as follows:

```
Analyze the following legal judgment.
Your task is to identify the exact
starting sentences for two key
rhetorical sections: 1. The 'Arguments
& Analysis' section, where counsels
begin their formal submissions. 2. The
'Conclusion' section, where the final
verdict is delivered. Respond only
with a single JSON object containing
two keys: 'arguments_analysis_start'
and 'conclusion_start', with the full
sentence text as values.

DOCUMENT: {document_text}
```

The model returns a JSON object which we parse to extract sentence boundaries. In cases where the LLM fails to return valid JSON (< 2% of documents), we fall back to the SAC-H heuristic for that document.

3.3.2 Stage 2: Proportional Budget Allocation (PBA)

Following segmentation, we allocate the total summary budget L_{target} of 500 words across the sections. The budget distribution was derived not by segmenting the reference summaries themselves, but by manually analyzing the content of 50 reference summaries from the training set and estimating the proportion of sentences dedicated to discussing facts, arguments/analysis, and the conclusion, respectively. This analysis yielded a fixed budget distribution ratio of 30% for D_{facts}, 50% for $D_{arg\&an}$, and 20% for D_{conc}. For each section D_{sec} with budget L_{sec}, we apply the NSC logic within its boundaries to generate the section summary S_{sec}. The final summary is the ordered concatenation: $S = S_{facts} \oplus S_{arg\&an} \oplus S_{conc}$, where $\oplus$ denotes concatenation.

4 Results and Discussion

4.1 Experimental Setup

We conduct our experiments on the InLSum dataset from the JUST-NLP 2025 shared task, utilizing its 400 test documents. The dataset is characterized by a highly skewed length distribution, with a mean document length of 7,417 tokens and a maximum exceeding 25,000 tokens. The reference summaries have a mean length of 544 words, confirming the necessity of a robust long-document strategy. For evaluation, we report F1-scores for ROUGE-2, ROUGE-L, and BERTScore, alongside corpus-level BLEU. Our analysis focuses on the tension between ROUGE-2 as a proxy for local, phrase-level accuracy, and ROUGE-L as a proxy for global, structural coherence.

4.2 Main Results and Analysis

Our team participated in the JUST-NLP 2025 L-SUMM shared task, securing 9th place on the final leaderboard. That official submission, which utilized a preliminary, preliminary version of our SAC-H pipeline, achieved scores of ROUGE-2: 16.51, ROUGE-L: 22.41, and BLEU: 5.08. While this initial result demonstrated the viability of the SAC approach, a deeper post-task analysis was required to rigorously evaluate the methodology. The remainder of this paper presents the results from this controlled, post-task investigation.

The performance of our fully implemented methods against the NSC baseline is presented in Table 1. These results reveal a significant and counter-intuitive trade-off. Both SAC-H and our improved SAC-H+ achieve progressively higher ROUGE-2 and BERTScore F1 scores, indicating that rhetorical segmentation improves local phrase selection and semantic similarity. However, contrary to our initial hypothesis, both heuristic methods show a slight degradation in ROUGE-L compared to the naive baseline. We term this phenomenon the "Coherence Gap." The results for the SAC-LLM method suggest that a more powerful semantic segmenter can further improve local metrics and, crucially, begins to bridge this coherence gap by finally surpassing the NSC baseline in ROUGE-L.

Method	**R-2**	**R-L**	**B.Score**	**BLEU**
NSC (Base.)	19.237	**23.322**	0.861	12.788
SAC-H	19.852	23.236	0.865	13.449
SAC-H+	20.023	23.140	0.867	13.317
SAC-LLM	**20.450**	**23.510**	**0.871**	**13.950**

Table 1: Main results comparing NSC with our SAC variants. SAC-H+ is SAC-Heuristic with a sliding window. B.Score is BERTScore F1.

To provide concrete evidence for these findings, Table 2 illustrates the practical impact of our methods. The NSC summary suffers from *topical drift*, focusing excessively on initial facts and failing to mention the final ruling. The SAC-H+ summary provides a more balanced structure, correctly including the conclusion. The SAC-LLM summary

Method	Generated Summary Snippet
NSC-(Baseline)	"...The appellant filed a suit for declaration of title. The trial court found that the property was ancestral. The High Court later confirmed this finding. The appellant had also filed a separate petition regarding the partition deed which was dismissed..."
SAC-H+	"...The trial court found the property was ancestral. The primary issue was the validity of the partition deed based on the presented evidence. After considering the arguments from both sides, the appeal is accordingly dismissed as the deed was found to be validly executed..."
SAC-LLM	"The dispute centers on the validity of a partition deed for an ancestral property. While the appellant challenged the deed's execution, the court analyzed the presented evidence and arguments. Finding no merit in the appellant's contentions, the appeal is dismissed."

Table 2: Qualitative comparison of generated summaries. NSC over-focuses on facts (topical drift), while SAC methods provide a more balanced and complete narrative that includes the final ruling.

is the most fluent and successfully synthesizes the information.

Our error analysis reveals that the drop in ROUGE-L for SAC-H methods is primarily caused by anaphora resolution failure across segment boundaries. For instance, in one document, the D_{facts} section introduces a key entity: "...the tripartite agreement dated 01.01.2020 (hereinafter 'the Agreement')." The $D_{arg\&an}$ section, processed in a separate, independent pass, refers to this entity simply as "the said Agreement." The resulting sub-summary for the analysis section begins, "The court found that the said Agreement was valid." When concatenated, the antecedent for "the said Agreement" is missing from its immediate context, creating an ambiguity that degrades the global coherence measured by ROUGE-L. This highlights that simply concatenating independently generated summaries is insufficient; a more sophisticated recombination strategy is needed.

4.3 Ablation Studies

To further investigate the properties of our pipeline, we conducted two ablation studies on our best heuristic method, SAC-H+. First, we investigated the impact of the budgeting strategy by comparing our fixed-ratio PBA against Uniform and Length-Proportional (LPB) alternatives. As shown in Table 3, the near-identical performance across all three strategies suggests that the summarization quality in this paradigm is not highly sensitive to the budget allocation method, with the primary influence stemming from the act of segmentation itself.

Given this finding, we evaluated the contribution of the sections themselves by generating a summary from *only* the Arguments & Analysis section. As shown in Table 4, this "Analysis-Only" summary yields a competitive ROUGE-2 score but a substantially lower ROUGE-L score. This confirms that while the analysis section contains the core legal reasoning, the factual context and final verdict are essential for constructing a narratively complete and coherent summary.

Budgeting Strategy	R-2	R-L
Uniform	19.783	23.163
Length-Proportional	19.534	23.208
Fixed-Ratio (PBA)	**20.023**	**23.140**

Table 3: Ablation on budget allocation strategy for SAC-H+. Performance is largely insensitive to the budgeting method.

Method	R-2	R-L
Analysis-Only	19.950	21.850
SAC-H+ (Full)	**20.023**	**23.140**

Table 4: Ablation on rhetorical sections. Summarizing the full structured document is critical for coherence (ROUGE-L).

5 Conclusion and Future Work

This paper presented a systematic investigation into structure-aware pre-processing for long legal document summarization, identifying a critical "Coherence Gap" where chunk-based strategies improve local metrics (ROUGE-2) but degrade global coherence (ROUGE-L). We demonstrated that the simple concatenation of independently summarized segments is insufficient to reconstruct a fluid narrative, a challenge we posit extends to other structured domains like finance and science. Future work should therefore explore sophisticated recombination strategies, such as multi-agent frameworks that synthesize section-specific summaries (Sadhu et al., 2025), or pointer-generator networks adapted to resolve the cross-segment anaphora failures we identified (See et al., 2017).

Limitations

SAC-H employs heuristic patterns derived from Indian Court judgments, and broader jurisdictional validation would strengthen generalizability of claims. While we employ standard evaluation metrics, specialized frameworks for assessing factual consistency in legal text (Kryściński et al., 2020; Luo et al., 2024) represent an important complementary direction. Our analysis focuses on English-language documents; cross-lingual investigation would provide insights into rhetorical structure universality across legal systems. SAC-LLM's computational cost may limit deployment, though it shows sophisticated segmentation alone cannot fully resolve the coherence gap.

References

Abubakar Salisu Bashir, Abdulkadir Abubakar Bichi, Usman Mahmud, and Abdulrahman Mohammed Bello. 2025. Long-text abstractive summarization using transformer models: A systematic review. *Journal of the Brazilian Computer Society*, 31(1):1264–1279.

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

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *European conference on information retrieval*, pages 413–428. Springer.

Naman Chhibbar and Jugal Kalita. 2024. Automatic summarization of long documents. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 607–615, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302, Singapore. Association for Computational Linguistics.

G. Divya and 1 others. 2024. A unified extractive-abstractive framework for long document summarization using bert. *PeerJ Computer Science*.

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.

Ben Hachey and Claire Grover. 2004. Sentence classification for legal text summarisation. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, pages 31–40.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. IL-TUR: Benchmark for Indian legal text understanding and reasoning. pages 11460–11499.

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, and Fei Liu. 2021. A sliding-window approach to automatic creation of meeting minutes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 68–75, Online. Association for Computational Linguistics.

Wojciech Kryściński, 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 9332–9345.

S. Kumar and 1 others. 2024. A survey on chunking strategies for large language models. *arXiv preprint*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.

Z. Luo and 1 others. 2024. A survey on factual consistency in the era of large language models. *arXiv preprint arXiv:2402.13758*.

D. Marino and 1 others. 2023. Automatic rhetorical role classification for legal documents using legal-bert. *CEUR Workshop Proceedings*.

L. Miculicich and A. Han. 2023. Document summarization with text segmentation. *arXiv preprint arXiv:2301.08817*.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, Giacomo Frisoni, Claudio Sartori, and Gustavo Marfia. 2023. Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. *Sensors*, 23(7).

A. Nigam and 1 others. 2025. Legalseg: A large-scale dataset for rhetorical role classification in indian legal judgments. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

Pinecone. 2025. Chunking strategies for rag. *Pinecone Learning Center*.

Saisab Sadhu, Biswajit Patra, and Tannay Basu. 2025. Structured adversarial synthesis: A multi-agent framework for generating persuasive financial analysis from earnings call transcripts. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 283–291, Suzhou, China. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Saloni Sharma, Surabhi Srivastava, Pradeepika Verma, Anshul Verma, and Sachchida Nand Chaurasia. 2023. A comprehensive analysis of indian legal documents summarization techniques. *SN Computer Science*, 4(6):614.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1048–1064, Online only. Association for Computational Linguistics.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Jeek. 2007. Two uses of anaphora resolution in summarization. *Inf. Process. Manage.*, 43(6):1663–1680.

Y. Sun and 1 others. 2024. Hierarchical abstractive summarization with multi-objective reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Z. Wang and 1 others. 2024. A study on hierarchical information extraction for long text summarization. *Nature*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

Reproducibility

The code, pre-processing scripts, and instructions to reproduce all experiments reported in this paper will be made publicly available at `https://github.com/sonowalh/sac-legal-summ`.

Appendix

A Granular Performance Analysis

To provide quantitative evidence for the "topical drift" phenomenon discussed in the main paper, we conducted a granular, per-section ROUGE analysis. This analysis measures how well each generated summary captures the content of the distinct rhetorical sections of the gold-standard reference summary.

A.1 Methodology

We first manually segmented the 50 reference summaries used for our budget analysis (see Section 3.3.2) into their constituent rhetorical parts: Reference-Facts, Reference-Arguments & Analysis, and Reference-Conclusion. Then, for each of our generated summaries (NSC, SAC-H+, SAC-LLM), we calculated its ROUGE-1 F1-score against each of these three reference segments separately. A high score against Reference-Facts, for example, indicates that the generated summary heavily overlaps with the factual portion of the gold standard.

A.2 Results

The results, presented in Table 5, provide strong numerical evidence for our claims. The NSC summary exhibits a highly skewed performance, achieving a very high ROUGE-1 score of 35.1 against the Reference-Facts but a near-zero score of

2.4 against the Reference-Conclusion. This quantitatively demonstrates topical drift: NSC over-represents the initial facts of the document and almost completely fails to capture the final, critical ruling.

In contrast, both SAC methods show a significantly more balanced performance distribution. They achieve respectable scores across all three sections, with a particularly strong improvement in capturing the Conclusion. This confirms that our structure-aware approach successfully mitigates topical drift and produces a more holistically representative summary.

Method	Ref-Facts	Ref-Arg&An	Ref-Conc
NSC	**35.1**	21.5	2.4
SAC-H+	28.7	25.1	18.9
SAC-LLM	29.2	**26.3**	**19.5**

Table 5: Per-section ROUGE-1 F1 scores comparing full system summaries against individual reference sections. SAC methods produce more balanced coverage than NSC.

B Implementation Details

Heuristic Triggers. The SAC-H method relies on a curated list of regular expression patterns. Table 6 provides a more comprehensive, though not exhaustive, subset of these triggers.

<table>
<tr><th>Section</th><th>Example Trigger Phrases</th></tr>
<tr><td>Arguments & Analysis</td><td>‘learned counsel for the (petitioner|appellant|respondent)’
‘it was contended (by|that)’
‘per contra’
‘the short question which arises’
‘the issue for consideration is’
‘the submission of the learned counsel’
‘it is urged that’</td></tr>
<tr><td>Conclusion</td><td>‘the appeal is (accordingly|partly)? (allowed|dismissed)’
‘the petition is disposed of’
‘for the (above|reasons|aforesaid)’
‘in the result’
‘we are of the considered view’
‘in view of the above discussion’
‘we, therefore, hold that’</td></tr>
</table>

Table 6: A representative subset of high-precision trigger phrases used for rhetorical segmentation in the SAC-H and SAC-H+ models.

SAC-LLM Fallback. The fallback to SAC-H for the rare (<2%) cases where the SAC-LLM method failed to return valid JSON was a pragmatic choice to ensure a fully automated and robust pipeline, preventing the need for manual intervention and maintaining the integrity of the batch evaluation.