STRONG – Structure Controllable Legal Opinion Summary Generation

Yang Zhong
University of Pittsburgh
Pittsburgh, PA
yaz118@pitt.edu

Diane Litman
University of Pittsburgh
Pittsburgh, PA
dlitman@pitt.edu

Abstract

We propose an approach for the structure controllable summarization of long legal opinions that considers the argument structure of the document. Our approach involves using predicted argument role information to guide the model in generating coherent summaries that follow a provided structure pattern. We demonstrate the effectiveness of our approach on a dataset of legal opinions and show that it outperforms several strong baselines with respect to ROUGE, BERTScore, and structure similarity.

1 Introduction

Discourse structure plays an essential role in text generation in domains ranging from news (Van Dijk, 2013) to peer-reviewed articles (Shen et al., 2022b). In the legal domain, it's equally important to draft a summary that can follow a blueprint (Xu et al., 2021). For instance, in Figure 1, given a long legal opinion with thousands of words as input, a legal expert organized the summary by making the argument clear in terms of the issues the decision addressed, the decision's conclusion, and the reasoning behind the decision.

While progress has been made in controllable generation, limited research has controlled discourse structure. Recently, Spangher et al. (2022) and Shen et al. (2022a) proposed approaches to generate sentences with discourse structure labels. However, no existing controllable generation work addresses the legal domain, where the argumentative structure is pivotal. While prior work in the legal field highlighted the significance of argumentative structure from the input (Elaraby and Litman, 2022), the potential for utilizing argument structure to guide text generation remains unexplored.

Based on a corpus analysis showing that experts use common patterns to summarize legal opinions (the most frequent one is shown in Figure 1), we develop a novel structure-prompting approach called STRONG (**S**tructure con**TR**ollable legal **O**pi**N**ion summary **G**eneration). STRONG is implemented using Longformer Encoder Decoder (Beltagy et al., 2020) coupled with automatically created structure prompts. Results demonstrate that STRONG outperforms summarization models without structure control and improves inference time over models with structure control from other domains. We make our models available at https://github.com/cs329yangzhong/STRONG.

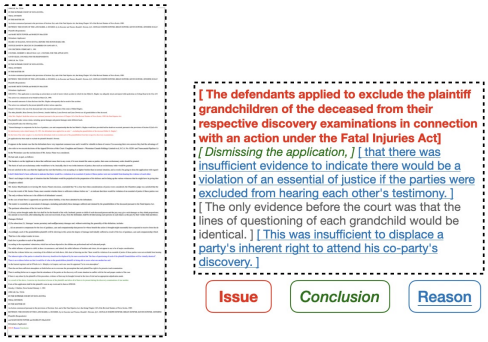

Figure 1: Example of a legal case opinion with its summary. The summary is annotated with oracle argument structure labels (one **Issue**, one *Conclusion*, and two Reasons). Presenting an issue followed by a conclusion and reasons is the dataset's most common normalized structure pattern (54%). Complete descriptions of patterns are in Appendix A.

2 Related Work

Prior work on controllable generation (Hu et al., 2017; Goyal and Durrett, 2020b; Dou et al., 2021; He et al., 2022) has focused on inner-sentence token-level attributes (e.g., syntactic structure) or full-text stylistic features (e.g., sentiment/topic). Recent research started looking at generating long texts adhering to discourse structures derived from news or article reviews (Ghazvininejad et al., 2022; Ji and Huang, 2021; Spangher et al., 2022; Shen et al., 2022b). Shen et al. (2022a) framed the task

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 431–448
November 1–4, 2023.

Split	Case/Summ pairs	Case len	Summ len	sents
No Manual Annotations				
Train	21794	3979.4	276.2	10.9
Valid	2724	4067.4	279.8	11.0
Test	2723	3899.9	278.8	10.9
Manual IRC Annotations				
1049-test	1049	3741.1	245.4	11.0

Table 1: Dataset statistics of CanLII. Case/Summary len is the text length in terms of the number of words, while sents is the sentence count per summary.

as a sentence-by-sentence generation, which led to a longer inference time compared to token generation baselines. *We explore structure control in legal opinions, which is challenging due to long input texts and argumentative discourse structures.*

In the legal domain, besides directly adopting the raw document-summary pairs into supervised training using abstractive summarization models such as BART (Lewis et al., 2020) and Longformer Encoder Decoder (LED) (Beltagy et al., 2020), Elaraby and Litman (2022) proposed highlighting the salient argumentative sentences in the inputs and training a model that is argument-aware. *We instead focus on improving argument structure adherence by exploiting the summaries' annotated discourse structures to create structure prompts rather than by manipulating the original articles.*

3 Dataset

We leverage the CanLII dataset of legal case opinions and human-written abstractive summaries.[1][2] It consists of 28,290 legal opinions and human-written summary pairs. For testing, we first leverage the annotated subset produced by Xu et al. (2021), including 1,049 pairs with manually annotated **IRC argument labels**: *Issues* (the legal questions addressed in the case), *Conclusions* (the court's decisions for the related issue), *Reasons* (text snippets illustrating the reasons for the court's decision) and *Non_IRC* (none of the above). We further split the remaining 27,241 unannotated pairs into 80/10/10 percent for model training, validation, and extra testing. Corpus statistics are in Table 1.

As introduced in §1 and Figure 1, legal experts devised different strategies to construct the summaries. We thus analyze the patterns of the IRC labels in the 1,049 annotated summaries. To comprehend the high-level structures better, we remove the Non_IRC tags and collapse adjacent text segments with the same tag into one. The most common "normalized" patterns are "Issue – Conclusion – Reason" (54%) and "Issue – Conclusion – Reason – Conclusion" (9%). Pie charts of the top normalized and original patterns are in Appendix A.

4 Method

Figure 2 illustrates our proposed STRONG approach. We start by extending the small-scale annotations to the larger dataset. Since we only have the 1,049 test set manually annotated with oracle summary argument labels, different from Elaraby and Litman (2022) who used a classifier on input sentences, we propose to train a sentence classifier on summary sentences (Stage 1) and then utilize it to predict silver labels for all unannotated summaries in Stage 2.[3] Our approach distinguishes itself from Shen et al. (2022b), which relied solely on manually annotated structure sequences, resulting in a smaller training set than our larger dataset with silver labels. In the next step of Stage 2, we introduce special marker tokens to guide the model in generating summaries following specified structure patterns. Specifically, we extract the argumentative "IRC" labels from summary sentences, concatenate them with split " | " tokens and prepend before the original input text, and connect them with a special marker "==>". This operationalizes the argument mining of salient information blueprint, providing better guidance for the model in generating legal summaries. That is, Stage 2 utilizes the predicted structure labels to fine-tune the LED model. Once the model has been trained, we generate summaries using different sets of structure labels for the two test sets during Stage 3 of the inference process.

5 Experimental Setup

We compare STRONG to two baselines. **NoStructure** uses the Longformer-Encoder-Decoder (LED) base model for generating summaries. The second baseline re-implements **SentBS** (Shen et al., 2022a)

[1]The data was obtained through an agreement with the Canadian Legal Information Institute (CanLII): https://www.canlii.org/en/

[2]The corpus is moderately abstractive: The overlap ratios for the 1/2/3-gram between the source document and the human-authored summaries stand at 89.7%, 62.0%, and 42.1%, respectively, which suggests a moderate level of abstractiveness of the dataset compared to others such as TL;DR (Völske et al., 2017). It can thus serves as a useful testbed for abstractive summarization.

[3]We include the model details in Appendix B.2.

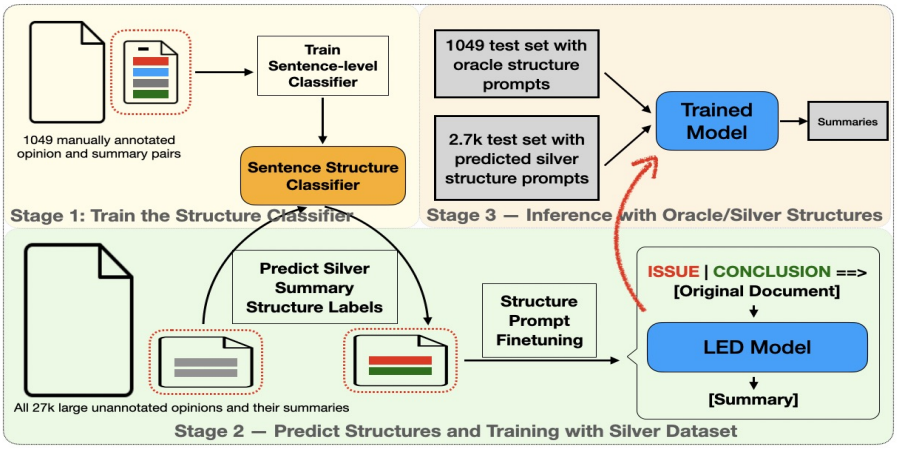

Figure 2: Illustration of our structure prompting approach (STRONG).

and is structure-aware. It uses a prompt-based backbone model to generate sentences, optimizing candidate selections based on the model likelihood and structure label probability. All implementation details are in Appendix B.

All experiments are evaluated using ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) F1 (Lin, 2004), BERTScore (BS) (Zhang et al., 2020), and structure similarity (*SS*) (Shen et al., 2022b). More details on the structure metric are in Appendix C.

6 Results and Analysis

6.1 Automatic Result

This section addresses two research questions: **RQ1**. Does STRONG improve summarization quality compared to baselines? **RQ2**. How do models compare in preserving structure? We then conduct analyses based on the observations and perform a small-scale human evaluation.

RQ1. Using the left results section of Table 2, we first compare STRONG with the NoStructure baseline on traditional ROUGE and BERTScore summarization metrics. For the 1049 test set, when the maximum generation output length is limited to 256 tokens, we observe that STRONG obtains an average of 2.1, 0.7, 2.1, and 0.2 improvements across ROUGE-1, 2, L, and BERTScore (rows 3 vs. 2), which are **significant** based on 95% confidence intervals. STRONG also outperformed the re-implemented SentBS baseline (rows 3 vs. 1). We also explored the impact of increasing the maximum output length to 512 tokens, based on the observation that oracle summaries tended to be longer (Table 1). Similar trends were seen when the maximum output length is increased to 512 tokens (rows 5 vs. 4), as well as when all analyses are repeated using the 2,723 silver set (rows 6-8, 9-10). This illustrates that the target structure information helps STRONG generate higher-quality summaries. Appendices D and E present examples and analysis to demonstrate model output differences in content coverage.

RQ2. In the 1049 test set, compared to the NoStructure model (row 2), the STRONG model (row 3) significantly improves the structure similarity scores by 0.03. While SentBS (row 1) outperforms both methods (rows 2/3), the tradeoff is increasing inference time (last column). In contrast, with the extended 512 generation length where we could not even run SentBS, STRONG obtained the best oracle test set performance in the table, with a margin of 0.1 compared to SentBS (rows 5 vs. 1). Albeit imperfect, on the silver test set where our IRC sentence classifier predicts the structure labels, STRONG also gains 0.1 improvements to NoStructure (rows 7 vs. 8, and 9 vs. 10), and now even surpasses SentBS (row 6 vs. 8) on structure similarity while again reducing inference time.

6.2 Length Control

The second to last column of Table 2 shows that STRONG generates the longest summaries, which may have impacted the above assessments. We thus force NoStructure and STRONG to continue generating tokens until reaching the same specified limit of {64, 128, 256, and 512} tokens.[4] Table 3

[4]The generation length of SentBS cannot be rigidly regulated, considering that it adheres to a sentence-by-sentence generation paradigm, and the inconsistencies in the length of structural prompts result in diverse outputs.

ID	Model	R-1	R-2	R-L	BS	SS	Avg Length	Infer. Time
				1049 Oracles				
Max output of 256 tokens								
1	SentBS	48.31	23.86	44.73	86.87	0.436	129.6	8.5 hours♦
2	NoStructure*	50.33	25.84	46.47	87.39	0.344	159.2	2.2 hours
3	STRONG*	52.47	26.54	48.57	87.63	0.372	186.3	2.5 hours
Max output of 512 tokens								
4	NoStructure	51.61	26.72	47.76	87.49	0.383	198.1	4.2 hours
5	STRONG*	**55.90**	**28.61**	**51.97**	**87.78**	**0.535**	263.0	4.3 hours
				2723 Silver Test Set				
Max output of 256 tokens								
6	SentBS	49.24	25.43	45.58	85.47	0.470	118.0	21.5 hours♦
7	NoStructure*	50.76	26.84	46.78	87.75	0.330	160.6	6.2 hours
8	STRONG*	52.84	27.90	48.73	87.97	0.493	179.3	6.3 hours
Max output of 512 tokens								
9	NoStructure	52.22	27.57	48.18	87.69	0.440	196.9	13.0 hours
10	STRONG*	**57.17**	**29.87**	**52.93**	**88.10**	**0.543**	255.9	13.1 hours

Table 2: Results of different models on the CanLII oracle and silver test sets. BS refers to BERTScore, SS means structure similarity, respectively. Models with * mean all results are statistically different from the previous row, based on 95% confidence intervals. All results are reported as an average of 3 runs initialized with random seeds. Best results are highlighted with **bold**, and best results under the 256 token settings are underlined. Rows 1 and 6 (with ♦) experiment with an RTX3090Ti card with larger memory, which will make the inference time faster than on the default cards, which are RTX5000s and used for all other experiments.

Model	Control Len.	R-1	R-2	R-L	BS
NoStructure	No	50.33	25.84	46.47	87.39
STRONG	No	52.47	26.54	48.57	87.63
NoStructure	Yes	50.74	25.91	47.07	87.17
STRONG	Yes	50.96	26.26	47.33	87.39

Table 3: Results of models when summary has a maximum (top) versus controlled (bottom) length of 256 tokens. Although STRONG still outperforms the baseline, the delta is reduced when the length is controlled.

shows the results for the 256 token limit,[5] and indicates that the Table 2 performance gap (repeated in the first two rows of Table 3) diminishes when the length is controlled (the last two rows). This suggests that the structural benefits of STRONG become less important when output length is fixed. However, controlled length can lead to incomplete generations (see an example in Appendix D.1), and STRONG can dynamically adjust and generate similar length summaries compared to the oracle when they can stop generation if needed. Additionally, for both NoStructure and STRONG, we observe a drop in ROUGE performance for extremely long summaries (512 tokens) compared to smaller output lengths (see Appendix D.1), likely because 512 tokens deviate from the distribution of human summarization lengths. We additionally experimented with another setup to adjust the minimum generation length of each model and with higher length penalties. These results are detailed in Table 9, located in Appendix D.2. We observed that our STRONG model outperformed the baseline and reinforced the notion that structural information plays a crucial role in guiding the model to produce summaries with the appropriate length and level of detail.

Model	SUMMAC$_{\text{Conv}}$
Max output of 256 tokens	
SentBS	0.660
NoStructure	0.663
STRONG	0.704*
Max output of 512 tokens	
NoStructure	0.658
STRONG	0.697*

Table 4: Results of the average factuality scores for models in Table 2 over the CanLII oracle test set. * means the result is significantly different from the previous row using paired t-test.

[5] An analysis of additional lengths is in Appendix D.1.

6.3 Factuality

To evaluate the factuality of generated text, we picked the SUMMAC$_{\text{CONV}}$ score from Laban et al. (2022), which utilizes the NLI model to detect summary inconsistencies and performs well on multiple factuality benchmarks (details in the original paper) compared to other metrics such as FactCC (Kryscinski et al., 2020) and DAE (Goyal and Durrett, 2020a). As shown in Table 4, our STRONG model obtains the highest scores, which means the highest consistency between document and generated summaries.

6.4 Human Evaluation

Human evaluation is under-explored for legal tasks, as it is labor-intensive due to long documents / summaries and requires evaluators with legal expertise (Jain et al., 2021). As a first step, we conducted a small-scale human evaluation using five legal decisions to assess the quality of summaries generated by all models in Table 2. Three legal experts were asked to evaluate the coherence of the generated texts and assess the coverage of argumentative components when compared to the oracle summaries crafted by the human CanLII experts.[6] The evaluator feedback indicated that longer summaries could potentially introduce more factual errors, and there was inconsistency in terms of fluency and readability, with mixed performance observed (one annotator reported issues in two cases). On the other hand, the advantage of controllable structure generation was more evident when generating longer summaries. In two out of five cases, the summaries generated by STRONG were preferred in the 512-length setting, while under the 256-length setting, only one STRONG-generated summary was favored.

7 Conclusion

We proposed the STRONG approach for improving the summarization of long legal opinions by providing target-side structure information. STRONG accepts different types of prompts and generates summaries accordingly. Experiments demonstrated that the content coverage, summary length, structure adherence, and inference time are all improved with STRONG compared to prior structure-control and no-structure baselines.

[6]We provide the evaluation details in Appendix F.

Limitations

Our research results are constrained by our dependence on a single dataset for experimentation as well as by computing resource limitations. While prior work demonstrated that the SentBS approach could obtain negligible performance drop with regard to automatic metrics such as ROUGE and BERTScore compared to a finetuning structure prompted baseline, our current experiment is hindered by extreme demand of GPU memories given the much longer legal input and large parameter searching space. We also demonstrate that the slowness of compared work is more severe when transferring the model to our tasks. Further experiments on more extensive setups of the prior baselines can be important for future work to verify the past work's conclusions. We recognize that our methodology relies on annotated data for structure labels, particularly when adapting to novel domains. In future research, we aim to investigate zero-shot learning techniques to enable structure classification without the necessity for annotations.

While our paper uses standard summarization metrics and a similarity measure particularly related to our focus on structure controllability, we do not yet extensively investigate how STRONG impacts factuality besides the SUMMAC$_{\text{CONV}}$ score (Laban et al., 2022). A recent study (Wan et al., 2023) demonstrates that improvements in factuality-related metrics come with the sacrifice of dropping automatic metrics such as ROUGE and BERTScore, while Min et al. (2023) harness the power of LLMs to evaluate the factuality of long-form text generation. Deviating from prior work (Zhong and Litman, 2022) that studies the extractive summarization task, we focused on the abstractive summarization, which has shown to surpass the performance of extractive methods by a noticeable margin, while both strategies introduce unfaithfulness (Zhang et al., 2023). Another limitation is that we only exploited the IRC structure representations due to the availability of oracle summary annotations. Exploring the use of structures based on other methods such as Lu et al. (2018) is a promising area for future work. Also, the automatic evaluation metrics may be deficient compared to human evaluations, thus unfaithfully representing the final quality of generated summaries compared to real legal experts. Moreover, in a real application, end users may propose and inquire about different out-

puts with self-designed structure prompts[7], which remains an open-ended challenge and may need human validation for future works.

Ethical Considerations

Using generated abstractive summary results from legal opinions remains a problem, as abstractive summarization models have been found to contain hallucinated artifacts that do not faithfully present the source texts (Kryscinski et al., 2019; Zhao et al., 2020). The generation results of our models may carry certain levels of non-factual information and need to be used with extra care. Similarly, CanLII has taken measures (i.e., blocking search indexing) to limit the disclosure of defendants' identities, while abstractive approaches may cause potential user information leakage.

Acknowledgements

This work is supported by the National Science Foundation under Grant No. 2040490 and by Amazon. We want to thank the members of the Pitt AI Fairness and Law Project members, the Pitt PETAL group, and anonymous reviewers for their valuable comments in improving this work.

References

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. 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.

Marjan Ghazvininejad, Vladimir Karpukhin, Vera Gor, and Asli Celikyilmaz. 2022. Discourse-aware soft prompting for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4570–4589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020a. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020b. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 238–252, Online. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. pages 5879–5915.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *ICML'17*, page 1587–1596. JMLR.org.

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2021. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388.

Haozhe Ji and Minlie Huang. 2021. DiscoDVT: Generating long text with discourse-aware discrete variational transformer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4224, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. 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.

[7]We provide an example of feeding different prompts to generate diverse summaries in Appendix E.1.

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.

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.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengdong Lu, Xianggen Liu, Haotian Cui, Yukun Yan, and Daqi Zheng. 2018. Object-oriented neural programming (OONP) for document understanding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2726, Melbourne, Australia. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation.

Chenhui Shen, Liying Cheng, Lidong Bing, Yang You, and Luo Si. 2022a. SentBS: Sentence-level beam search for controllable summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10256–10265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022b. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535, Dublin, Ireland. Association for Computational Linguistics.

Alexander Spangher, Yao Ming, Xinyu Hua, and Nanyun Peng. 2022. Sequentially controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6848–6866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Teun A Van Dijk. 2013. *News as discourse*. Routledge.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Shiyue Zhang, David Wan, and Mohit Bansal. 2023. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. *ACL*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*. Association for Computing Machinery.

Yang Zhong and Diane Litman. 2022. Computing and exploiting document structure to improve unsupervised extractive summarization of legal case decisions. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 322–337, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A IRC Structure Patterns

We report the distribution of different structure patterns with the normalized version (we remove the neighboring duplicated labels and ignore the Non_IRCs for better structure presentation) in Figure 3. We observe that most 1049 test summaries are annotated in an Issue – Conclusion – Reasoning pattern, while the remaining have different reordering of latter patterns. Legal experts sometimes employ the "Conclusion then Reasoning" pattern (3.6%) to strengthen the validity of the case summary. We found 54 distinct normalized structure patterns without considering the Non_IRCs and varying numbers of neighboring sentences. This suggests that legal experts employed diverse strategies to construct the summaries and confirms the importance of structure modeling in text generation tasks. Regarding the original patterns (excluding Non_IRCs), as shown in Figure 4, the numbers of Issue and Reasoning sentences varied.

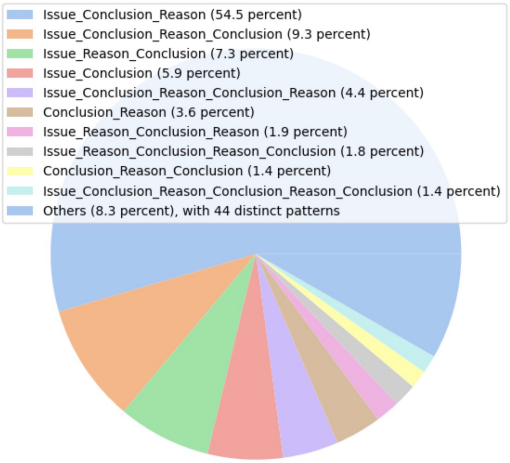

Figure 3: Pattern distribution of normalized summary structures, here we exclude the Non-IRC labels.

B Implementation Details

All of our BART-based experiments and the sentence classification model are conducted on Quadro RTX 5000 GPUs, each with 16 GB RAM. For SentBS models, we adopted the authors' original codebase workflow[8] and reimplemented it on an RTX 3090Ti GPU to satisfy the minimum RAM requirements.

[8] https://github.com/Shen-Chenhui/SentBS

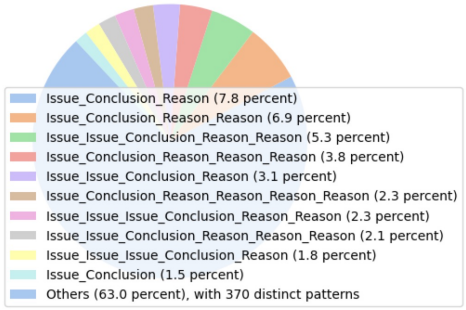

Figure 4: Pattern distribution of summary structures, here we exclude the Non-IRC labels.

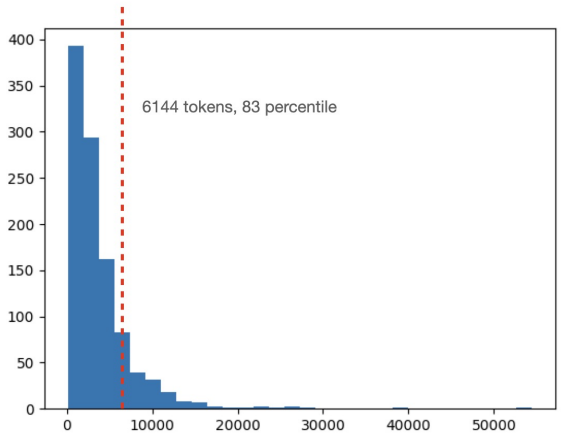

Figure 5: Input case length distribution of the 1049 test set, for models truncated at 6144 tokens, we retain 83 percent complete inputs.

B.1 NoStructure and STRONG model

All models are implemented with the Huggingface library (Wolf et al., 2020) using PyTorch, initialized with the *"allenai/led-base-16384"* checkpoint[9]. We train all our models with the same learning rate of 2e−5. We train the models for 16k steps, using the gradient step of 4, batch size of 1, and save the best checkpoints at every 1,000 steps, based on the ROUGE-2 F1 score of the validation set evaluations. Each model is trained with three randomized seeds, and we report the final averaged results. For training summarization models, we set the min/maximum inference summary length to 64/256 tokens. We employed beam-search with a beam size of 4 for all experiments. We additionally experimented with 512 output lengths in the main results. We truncate the input length to 6,144 tokens for the LED-base model due to our GPU

[9] https://huggingface.co/allenai/led-base-16384/tree/main

Data Split	I-F_1	R-F_1	C-F_1	Non-F_1	Macro F_1
Valid	76.7	66.3	76.7	76.0	73.8
Test	75.3	71.8	81.0	76.7	75.9

Table 5: IRC label classifier performance on the 1049 subset's validation and test split.

	R-1	R-2	R-L	Infer. time	Avg length
sentence-ctrl	48.31	23.86	44.73	8.5 hours	129.6
segment-ctrl	42.79	21.56	39.59	6 hours	77.7

Table 6: SentBS results with different structure sequences.

limitation, and analyze the effects of contents truncations.[10] For inference, we do a batch decoding with a batch size of 5 and report the total inference time accordingly.

B.2 IRC Classifier Training

Our argument role (IRC) classifier leverages a fine-tuned *legalBERT* (Zheng et al., 2021) model due to its performance gain compared to other contextualized models such as BERT (Devlin et al., 2019) and ROBERTa (Liu et al., 2019) as shown in Elaraby and Litman (2022) to predict sentence IRC labels as a four-way classification task. We implemented the model with the PyTorch Lightning framework[11]. We split sentences from the 1049 annotated summaries into a 80/10/10 randomized setting for train, validation, and testing. For model training, we set the learning rate of 2e−5, training for 15 epochs, and leveraged the validation loss for early stopping criteria with a patience of 5. The final prediction macro-F1 is 0.7586. The detailed sentence classifier result is shown in Table 5.

B.3 SentBS Re-Implementation

The original SentBS (Shen et al., 2022a) approach is implemented with a backbone of the BART-large (Lewis et al., 2020) model and using a V100 Graphic Card with 32GB memory. We first replaced the BART-large backbone with our trained LED models. Due to the limitation of GPU memory, the model failed to load on our prior RTX 5000 GPUs with the basic setting of beam size of 2. We instead ran the model on a GTX 3090Ti card with 24 GB memory, inference with the SentBS's "beam search + nucleus sampling" option, generation size of 4, beam size of 2, a top-p ratio at 0.9, and the maximum decoding length of 256 tokens. All other parameters are consistent with the original article experiment. Besides the sentence label searching, we additionally experiment with the segment-ctrl setup, where the target summary labels are de-duplicated to spans with non-repeated IRC labels. The results are shown in Table 6. We tested the model's performance on the original MReD dataset, which gives 34.77/9.69/30.99 regarding ROUGE scores, which is comparable to the original paper's result 34.61/9.96/30.87 with our evaluation script.

C Structure Similarity Evaluations

As mentioned in §5, we adopted a metric from the human evaluation introduced in Shen et al. (2022b) to measure the structure-similarity between a system output summary and a given oracle summary with the oracle structure prompt. In our actual implementation, the similarity score is computed by

$$1 - (\frac{minimum\_edit\_distance(S_i, O_i)}{max(len(S_i), len(O_i))}$$

where the edit distance is computed as the Levenshtein Distance, with equal penalties for replace, insert, and delete operations. We report the average similarity score of the test sets in the table results. Given that the sentence classification model can make wrong predictions, we estimate an upper bound by making predictions of the human-written summary sentences, which resulted in 0.781 for the original similarity score. Albeit not perfect, we can still assume that a generation model performs better on the structure-controlled generation task if the computed similarity becomes higher.

[10] We plot the length distribution of input documents in Figure 5.

[11] https://github.com/Lightning-AI/lightning

D More Analysis on the Generation

D.1 Controlled Length

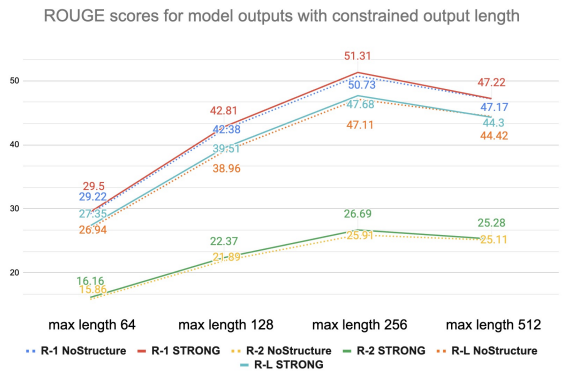

Figure 6: ROUGE scores for NoStructure and STRONG models with 64, 128, 256, and 512 output token limits.

We compare the ROUGE scores between the NoStructure and STRONG models and visualize the results in Figure 6. The findings suggest that the performance gap between the models diminishes, indicating that the structural benefits of the STRONG model for summary organization become less significant when the output length is fixed. Additionally, we observed a drop in performance for extremely long summaries (512 tokens) as they deviated from the distribution of human summarization lengths. The BERTScore performance is shown in Table 7, where we observe a similar trend.

However, controlled length can lead to the incomplete generation problem, as the model can not stop generation until it hits the desired token limit. As shown in Table 8, models obtain incomplete last sentences under the controlled length setting.

D.2 Controlled Min Length

We set the minimum length parameter of the generation to (64, 128, 256, and 512) and fixed the maximum length at 512. We modified the length penalty to 2.0, aiming to prompt the model to generate longer sequences. Table 9 indicates that our approach yields summaries with higher ROUGE scores when a larger length penalty is applied. This positive impact remains consistent even when we set the minimum length to less than 256 tokens. These findings reinforce the notion that structural information plays a crucial role in guiding the model to produce summaries with the appropriate length and level of detail. Interestingly, in the extreme scenario where we set the minimum tokens to 512, both models perform similarly.

D.3 Complete ROUGE scores

To evaluate the advantages brought by the proposed methods, alongside diagnosing the effects of augmenting the maximum generation length, we report the complete ROUGE scores of the models on the 1049 test set in Table 10. Initial observations highlight that the incorporation of structural information fosters enhancements in ROUGE recall scores, despite inducing a slight decrement in precision (as evidenced in row 2/3 and row 4/5). Additionally, the expansion of maximum output length significantly boosts the ROUGE recall, which can be attributed to the coverage of more n-grams. However, a corresponding decline in the precision score has been observed. This observation echoed with the preliminary human evaluation, which suggested that the longer outputs occasionally encompassed with higher error rate of contents, thus having lower quality.

Length	noStructure			STRONG		
	BS - P	**BS - R**	**BS - F1**	**BS - P**	**BS - R**	**BS - F1**
64	89.08	83.36	86.09	89.22	83.38	86.17
128	88.27	85.64	86.91	88.47	85.67	87.02
256	86.79	87.50	87.17	87.04	87.80	87.39
512	84.56	88.49	86.46	84.62	88.86	86.66

Table 7: Evaluation of models under Controlled Length, BS - P, BS - R, and BS - F1 denote BERTScore for Precision, Recall, and F1-Score, respectively. The table presents the evaluation results of models under different controlled lengths. There still exists difference between the two models, while overall the 512 length generation becomes worse.

Model	Generated Summary
Reference	The appellant was convicted of indecent assault against two young girls. He appealed on five grounds related to a substantial conflict in the evidence. Dismissing the appeal, that there was no error on the part of the trial judge in weighing the evidence.
NoStructure **max**	The appellant was convicted of two counts of indecent assault against two girls, aged 13 and 16. He was sentenced to nine months imprisonment, to be followed by two years probation. The appellant appealed.Dismissing the appeal, that there was no error on the part of the trial judge in conducting the trial or in weighing the evidence. After carefully reviewing the evidence, the verdict was not unreasonable or not supported by the evidence and the appeal was dismissed.
NoStructure **controlled**	The appellant was convicted of two counts of indecent assault against two girls, aged 13 and 17, respectively. He was sentenced to nine months imprisonment with respect to the first assault, followed by two years probation. The appellant appealed.Dismissing the appeal, that there was no error on the part of the trial judge in conducting the trial or in weighing the evidence. ..... There was a substantial conflict in the evidence as to the appellant's guilt, **and he had been sentenced to 9 months imprisonment for the assault on the complainant, to**
STRONG **max**	The appellant was convicted of two counts of indecent assault against two young girls. He was sentenced to nine months imprisonment with respect to the first assault, to be followed by two years probation, and one month consecutive for the second assault. The appellant appealed.Dismissing the appeal, that there was no error on the part of the trial judge in conducting the trial or in weighing the evidence. After carefully reviewing the evidence, the court could not say that the verdict was unreasonable or not supported by the evidence and the appeal was dismissed.
STRONG **controlled**	The appellant was convicted of two counts of indecent assault against two girls. He was sentenced to nine months imprisonment with respect to the first count and two years probation on the second count. The appellant appealed both convictions.Dismissing the appeal, that there was no error on the part of the trial judge in conducting the trial or in weighing the evidence. ..... **as the evidence did not support the appellant's contention that the assault was committed in bad faith and that the appellant had committed the second offence in good faith and in**

Table 8: A sample of 256 token generation for NoStructure and STRONG models under the max and control length settings. **Bold** sentences are incomplete under the controlled length setting.

Length	noStructure			STRONG		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
64	52.00	26.82	48.19	55.68	28.30	51.74
128	52.42	26.97	48.61	55.51	28.22	51.62
256	52.30	26.92	48.73	53.88	27.65	50.22
512	47.09	24.98	44.32	46.96	24.95	44.04

Table 9: Evaluation of models under Controlled Minimum Length. The table presents the evaluation results of models under different controlled minimum lengths. A difference still exists between the two models, while overall, the 512 length generation becomes worse.

ID	Model	R-1 Precision	R-1 Recall	R-1 F1	R-2 Precision	R-2 Recall	R-2 F1	R-L Precision	R-L Recall	R-L F1
1049 Oracles										
Max output of 256 token										
1	SentBS	59.93	44.93	48.27	29.75	22.10	23.80	55.65	41.45	44.67
2	NoStructure*	60.45	47.99	50.15	31.31	24.31	25.65	56.04	44.17	46.33
3	STRONG*	58.84	52.47	52.62	30.38	26.93	27.04	54.67	48.40	48.70
Max output of 512 token										
4	NoStructure	57.88	53.67	51.99	30.14	27.53	26.85	53.76	49.62	48.19
5	STRONG*	53.33	61.99	55.74	27.13	31.52	28.33	49.61	57.53	51.80

Table 10: The Complete ROUGE results for various models on the CanLII 1049 oracle dataset.

E Examples of Different System Outputs

E.1 Different Prompts' Effects

In Table 11, we generate multiple summaries according to different prompts using the best-performing STRONG method and set the maximum length of generation at 512 tokens. We find that the outputs follow the structure prompts to a certain degree. For instance, Variant 1 quickly jumped to the reasoning parts after the first two sentences, while Variant 2 started with multiple clear conclusion sentences on the court's decision and the main issues.

E.2 Sample Outputs

In Table 12, we show the examples for different methods under the 256 token max generation limit. We further ask three legal experts to rate the different outputs and analyze on the coverage of argumentative roles. We find that SentBS does a good job of stating an issue, but never reaches the conclusion. The NoStructure – 256 model fails to give a good statement of the issues, and our STRONG – 256 produces a more coherent and clear presentation. We additionally include the 512-token version of NoStructure and STRONG outputs in Table 13. Compared to the shorter NoStructure output that does not clearly state the issue, and it also doesn't reveal how the issue came out, the legal expert reported that the STRONG - 512 version is very clear and comprehensive. He also raised some concerns about the privacy problem of leaking the decedent's full name.

F Human Evaluation Details

We conducted evaluations with a total of three legal experts, all of whom hold a J.D. degree and possess a minimum of four years of experience in providing professional legal services. The experts were assigned five randomly sampled legal cases, each accompanied by the oracle reference summary, as well as the generated outputs from the following five models: (1) SentBS with a length of 256 tokens, (2) NoStructure with a length of 256 tokens, (3) STRONG with a length of 256 tokens, (4) NoStructure with a length of 512 tokens, and (5) STRONG with a length of 512 tokens. The experts were presented with the reference summary and all five system outputs in the same row of an Excel file. They were then asked to provide reflections on the faithfulness and coherence of each system output while considering the inclusion of essential argument roles components such as Issue, Reason, and Conclusion compared to the reference summary. Given that the instruction does not specifically inquire about the ranking nor ask evaluators to provide numerical scores, the primary author instead offers an interpretation of the free-text reflections by conducting comparative analyses across various outputs and allocating a relative ranking. We release all reflections for further studies in https://github.com/cs329yangzhong/STRONG. Table 14 shows one example of evaluators' reflections on a case, and Table 15 shows the author's ranking interpretation.

Prompt	Summary
Human Written Summary	
Issue | Conclusion | Conclusion | Reason	At issue was the interim custody of the child and child support. HELD: Interim custody was given to the petitioner with access to the respondent. The respondent was ordered to pay interim child support in the amount of $193 per month. On an interim custody application, the Court must take into account the status quo.
Generations	
Issue | Conclusion | Conclusion | Reason	At issue was the interim custody of the child and child support. The parties began cohabitating in May 1998 and separated in April 2000.HELD: The petitioner was awarded interim custody with reasonable access including every second weekend, reasonable telephone access and reasonable access during holiday periods. Both parents were ordered to complete the Parenting After Separation course. There was an unhealthy degree of constant bickering.
Variant 1: Issue | Conclusion | Reason | Reason | Reason	The petitioner was awarded interim custody with reasonable access including every second weekend, reasonable telephone access and reasonable access during holiday periods. Both parents were ordered to complete the Parenting After Separation course. The parties raised in their material the issue of each other's lack of cooperation and communication. While the Court was not in position to decide whether the accusations of each party were substantiated, it could say with relative certainty that there was an unhealthy degree of constant bickering. It was evident that as long as the parties work on establishing better communication, the child would benefit from contact with both parents with minimal disruption.
Variant 2: Conclusion | Issue | Issue | Reason | Reason	Interim custody of the child was awarded to the mother. The father was ordered to pay interim child support of $193 per month. Both parents were ordered to complete the Parenting After Separation course.HELD: The mother was awarded interim custody with reasonable access including every second weekend, reasonable telephone access and reasonable access during holiday periods. Interim child support was ordered in the amount of $125 per month and the father was to pay his proportionate share of childcare costs. There was an unhealthy degree of constant bickering. It was evident that as long as the parties worked on establishing better communication, the child would benefit from contact with both parents with minimal disruption.

Table 11: An example of the difference between generated summaries based on different prompts using our best structure prompt model STRONG. The original legal decision id is 2003skqb487.

ORACLE Statutes – Interpretation – Limitation of Civil Rights Act, Section 2 The plaintiff bank applied by notice of application for a deficiency judgment against the defendant pertaining to the non-purchase money portion of the outstanding indebtedness to the plaintiff pursuant to her mortgage and to some other costs related to the mortgage. The original mortgage executed by the plaintiff was for $191,000 of which $176,000 was advanced to purchase her condominium and $15,000, eight per cent of the total, was advanced for other purchases. The property sold for $156,000 pursuant to a judicial sale. The sale proceeds were distributed to pay outstanding property taxes, commission, and the principal and interest due under the mortgage in the amount of $141,945 and the net sale proceeds were to be distributed pro rata. As at November 1, 2017, the principal and interest owing on the non-purchase money portion of the mortgage was $21,250. The plaintiff requested judgment for the deficiency in the amount of $9,900 with interest to the date of judgment. It claimed it was entitled to solicitor-client costs and other costs accrued by it under the mortgage. The issue was whether such costs should be subject to the pro rata calculation for the nonpurchase portion of the mortgage, that is, 8 percent of the total costs incurred. HELD: The application was granted. The court found that the plaintiff was entitled to a deficiency judgment for principal and interest owed under the mortgage in the amount requested. It was also entitled to judgment against the mortgagee for eight percent of the solicitor-client costs, condominium fees and property management charges.
SentBS Statutes – Interpretation – Condominium Property Act, 1993, Section 63 The plaintiff bank applied by notice of application for a deficiency judgment against the defendant, Partridge, pertaining to the non-purchase money portion of the outstanding indebtedness to the plaintiff pursuant to her mortgage. Partridge had originally executed a mortgage for $191,000, of which $176,000 was advanced to purchase her condominium unit and $15,000 for other purchases. The non-Purchase money amount of the loan advanced and secured by the mortgage was 8RBC applied for solicitor-client costs and other costs accrued by RBC under the mortgage. The issue was whether the pro rata distribution applied solely to the indebtedness of principal and accrued interest, or applied to all costs of outstanding mortgage, including solicitor and client costs permitted under the mortgages. In its application, RBC had calculated the proceeds from the judicial sale that must be distributed pro ratus between the purchase money and non-paid balances owing, resulting in the deficiency of principal. For the deficiency amount to be set at the rate of 2.
NoStructure – 256 The petitioner was awarded interim custody with reasonable access including every second weekend, reasonable telephone access and reasonable access during holiday periods. Both parents were ordered to complete the Parenting After Separation course. The parties raised in their material the issue of each other's lack of cooperation and communication. While the Court was not in position to decide whether the accusations of each party were substantiated, it could say with relative certainty that there was an unhealthy degree of constant bickering. It was evident that as long as the parties work on establishing better communication, the child would benefit from contact with both parents with minimal disruption.
STRONG – 256 Mortgage – Foreclosure – Determination of InterestThe plaintiff bank applied by notice of application for a deficiency judgment against the defendant, Partridge, pertaining to the non-purchase money portion of the outstanding indebtedness to it pursuant to her mortgage. The plaintiff also applied to have its lawyers' costs assessed on solicitor-client basis. Partridge had originally executed a mortgage for $191,000, of which $176,000 was advanced to purchase her condominium unit and $15,000 for other purchases. Thus, the mortgage was 8 percent of the total mortgage.HELD: The plaintiff's application was granted and the plaintiff was granted judgment against Partridge in the amount of $9,901.23 with interest from November 1, 2017, to the date of judgment at the rate of 2.79 per cent per annum. It was entitled to solicitor and other costs accrued by the plaintiff under its mortgage.

Table 12: Different models' outputs for the legal decision (id: 5_2018skqb216). The structure prompt is "Non_IRC | Non_IRC | Non_IRC | Non_IRC | Non_IRC | Issue | Issue | Issue | Conclusion | Conclusion | Conclusion ".

NoStructure – 512 Mortgages – Foreclosure – DischargeThe plaintiff applied by notice of application for a deficiency judgment against the defendant pertaining to the non-purchase money portion of the outstanding indebtedness to the plaintiff pursuant to her mortgage. The plaintiff also applied to have its lawyers' costs assessed on solicitor-client basis.HELD: The plaintiff was granted judgment in the amount of $9,901.23 with interest from November 1, 2017 to the date of judgment at the rate of 2.79 per cent per annum for the deficiency of principal and interest owed under the mortgage.
STRONG – 512 Mortgage – Foreclosure – Pro RataDebtor and Creditor – Mortgage – Non-purchase MoneyThe plaintiff, Royal Bank of Canada, applied by notice of application for a deficiency judgment against the defendant, Doreen Pearl Partridge, pertaining to the non-payment of outstanding indebtedness to the plaintiff pursuant to her mortgage. The property sold pursuant to an Order Nisi for Sale by Real Estate Listing for $156,000 was distributed to pay outstanding property taxes, real estate commission and the principal and interest due under the mortgage in the sum of $141,945.36. At issue was whether the pro rata distribution applied solely to the indebtedness of principal and accrued interest, or applied to all costs of the outstanding mortgage, including solicitor-client costs permitted under it. Partridge had originally executed a mortgage for $191,000 of which $176,000 advanced to purchase her condominium unit and $15,000 for other purchases. Thus, the mortgage was 8 percent of the total mortgage.HELD: The plaintiff was granted judgment against Partridge in the amount of $9,901.23 with interest from November 1, 2017 to the date of judgment at the rate of 2.79 per cent per annum. It was entitled to solicitor and client costs and other costs accrued by the plaintiff under its mortgage, that is, 8 per cent of its total outstanding mortgage costs incurred. Section 63 of The Condominium Property Act, 1993 allows the condominium corporation to register a lien against the title of the unit for unpaid contributions to the common expense fund or the reserve fund. Secondly, the plaintiff claimed $1,461.92 for its payment of property management charges for securing and caring for the property, appraisal fee and utilities. These charges were permitted by s. 8(1) of the Limitation of Civil Rights Act (LCRA) and any inspections and administration fees had not been claimed by RBC. Further, the property management charge was recoverable under the terms of the mortgage.

Table 13: NoStructure and Strong models' outputs for the legal decision (id: 5_2018skqb216) under 512 max length generations. The structure prompt is "Non_IRC | Non_IRC | Non_IRC | Non_IRC | Non_IRC | Issue | Issue | Issue | Conclusion | Conclusion | Conclusion ".

Model Output	Annotator Reflection
SentBS	Does a good job stating an issue, but never reaches the conclusion.
NoStructure – 256	No good statement of the issue, but maybe readers could infer the issue based on the conclusion "It was entitled to solicitor and other costs accrued by the plaintiff under its mortgage." The interest payment isn't important enough to be in the summary
STRONG – 256	Fairly coherent, but it's not totally clear the dispute is about "solicitor and client costs permitted under the mortgages.
NoStructure – 512	It's not very clear about the issue, and it also doesn't reveal how the issue came out.
STRONG – 512	This is very clear. Using the defendant's full name might be a privacy problem. I also wonder if there's a copyright problem with using the subject classification system at the start of the summary. It looks like it's from the Law Society of Saskatchewan.

Table 14: A sample of the human evaluation on different model outputs. It corresponds to Annotator 3's reflection for the second legal decision summary group in Table 15.

Annotator	**SentBS**	**NoStructure-256**	**STRONG-256**	**NoStructure-512**	**STRONG-512**
Anno. 1	5	1	3	1	4 (too detailed)
Anno. 2			N/A		
Anno. 3	3 (fluency problem)	1	3 (same as sentbs)	2	5
Anno. 1	5 (no conclusion)	4 (no issue)	3	2	1 (fairly clear)
Anno. 2	5 (lack conclusion)	2	4 (factual errors)	2 (lack conclusion)	1 (very good)
Anno. 3	5 (never concludes)	3 (good)	3	2 (was nice, but ...)	1 (great)
Anno. 1			N/A		
Anno. 2	5 (wrong issue)	1	3 (too many details)	4 (erratic contents)	2 (too detailed)
Anno. 3	5 (very confusing)	1	3 (bad grammar)	4 (not reliable)	2
Anno. 1	3	3	5	1	2
Anno. 2	4 (no I, C)	4 (same to SentBS)	3 (no issue)	1	2 (fairly good)
Anno. 3			N/A		
Anno. 1			N/A		
Anno. 2	2 (not good)	2 (same to SentBS)	2 (same to SentBS)	1	5 (bad summary)
Anno. 3	3 (generally good)	3	2	1	5 (bad summary)

Table 15: The inferred rankings of different system outputs, determined based on human reflections over five legal decision summaries. Some annotators did not annotate a specific summary, and the row is represented by "N/A".