Adding Argumentation into Human Evaluation of Long Document Abstractive Summarization: A Case Study on Legal Opinions

Mohamed Elaraby†◇, Huihui Xu◇⋆, Morgan Gray◇⋆, Kevin Ashley◇⋆, Diane Litman†◇⋆
† Department of Computer Science, School of Computing and Information
◇ Learning Research and Development Center
⋆ Intelligent Systems Program, School of Computing and Information
University of Pittsburgh, Pittsburgh, PA USA
{mse30, hux16, mag454, ashley, dlitman}@pitt.edu

Abstract

Human evaluation remains the gold standard for assessing abstractive summarization. However, current practices often prioritize constructing evaluation guidelines for fluency, coherence, and factual accuracy, overlooking other critical dimensions. In this paper, we investigate *argument coverage* in abstractive summarization by focusing on long legal opinions, where summaries must effectively encapsulate the document's argumentative nature. We introduce a set of human-evaluation guidelines to evaluate generated summaries based on argumentative coverage. These guidelines enable us to assess three distinct summarization models, studying the influence of including argument roles in summarization. Furthermore, we utilize these evaluation scores to benchmark automatic summarization metrics against argument coverage, providing insights into the effectiveness of automated evaluation methods.

Keywords: Summarization, Human Evaluation, Legal Summarization

1. Introduction

Human evaluation remains the best practice for evaluating generated summaries (Kryscinski et al., 2019; Fabbri et al., 2021), although conducting such evaluations can be laborious and costly, particularly when dealing with longform summaries exceeding 150 words (Krishna et al., 2023; Karpinska et al., 2021; Clark et al., 2021; Goyal et al., 2022b). Consequently, most longform summarization research shies away from conducting human evaluation (Krishna et al., 2023). While recent efforts have attempted to tackle this issue by standardizing the evaluation process with a focus on the factual accuracy dimension of the generated summaries (Krishna et al., 2023; Min et al., 2023) or coherence (Goyal et al., 2022b), none have adequately accounted for the unique requirements of the domain, which may entail additional dimensions.

In this paper, we propose the integration of a new dimension, **argument coverage**, into the human evaluation of abstractive summarization. We define *argument coverage* as the ability of the generated summary to adequately include argument components from the source document. Our focus lies on *long legal opinions*, a type of legal document mainly concerned with court decisions and characterized by intricate implicit argument structures dispersed throughout lengthy texts (greater than 4000 words on average) (Xu et al., 2021; Elaraby and Litman, 2022; Elaraby et al., 2023; Zhong and Litman, 2023). The summaries are mostly considered longform summaries (greater than 200 words on average), Additionally, long legal opinions are composed of nuanced legal terminologies, necessitating legal experts for evaluation, which adds to the overall complexity of the task.

To address these research complexities, we make the following contributions: (1) We develop comprehensive human evaluation guidelines tailored for assessing argument coverage in generated abstractive summaries of long legal opinions. (2) We conduct a benchmarking study involving three existing systems, leveraging the introduced guidelines. This study aims to assess whether summarization models incorporating argument components achieve higher ratings of argument coverage compared to those that do not. (3) We assess the performance of automatic summarization metrics recently used in legal opinion summarization against human ratings, aiming to determine whether existing metrics adequately capture the variability in argument coverage within the generated summaries.

2. Related Work

Evaluating automatically generated summaries presents challenges such as scalability issues and low annotator agreement (Liu et al., 2023). These challenges are exacerbated when dealing with longform summaries, as assessing extended lengths inherently involves subjectivity (Karpinska et al., 2021). A comprehensive study by Krishna et al. (2023) revealed that 63% of research papers in longform summarization lack human evaluation. To

Proceedings of the HumEval @LREC-COLING 2024, pages 28–35
Turin, Italy. May 21st, 2024.

address this gap, they proposed guidelines for evaluating the factuality of longform summaries. Additionally, Min et al. (2023) introduced the FactScore metric to assess the factuality of long-generated summaries (biographies), breaking down factuality into atomic facts for comparison against ground truth. Another framework by Chang et al. (2023) focuses on assessing coherence in book-length summaries by leveraging Large Language Model evaluation capabilities. *However, there is limited work addressing evaluation methods for legal documents, which often produce longform summaries.*

In the pursuit of evaluating generated legal summaries, Mullick et al. (2022) undertook a human assessment focusing on the relevance and readability of legal summaries. Similarly, Salaün et al. (2022) conducted a human evaluation to assess the fluency and adequacy of legal summaries. Xu and Ashley (2023) had a legal expert evaluator who indirectly assesses the information quality of legal summaries by evaluating the quality of generated question-answer pairs. *In this study, human evaluators directly evaluated the legal argument coverage in generated legal summaries.*

In efforts to benchmark automatic metrics against human evaluations, Fabbri et al. (2021) conducted a benchmarking study on automatic summaries generated from 23 summarization models, sampled from the CNN-DailyMail dataset (Hermann et al., 2015). They evaluated these summaries using 14 distinct automatic summarization metrics across dimensions such as factual consistency, coherence, fluency, and relevance. Building upon this work, Liu et al. (2023) expanded the evaluation framework to include Atomic Content Units (ACUs), which are fine-grained semantic units enabling high inter-annotator agreement. These new evaluation scores were used to augment benchmark summaries, including those from the news domain (CNN-DailyMail and Xsum (Narayan et al., 2018)) and the dialogue domain (SamSum (Gliwa et al., 2019)), against automatic metrics. *In our study, we focus on benchmarking automatic metrics used in legal opinion summarization against human evaluation scores for argument coverage.*

3. Dataset for Evaluation

In this analysis, we utilized a subset of the **CanLII** dataset [1], consisting of 1049 cases annotated for argument roles types and summarization (Xu et al., 2021). The input legal opinions in this subset have mean and maximum lengths of 4375 and 62786 words, respectively, while the annotated summaries have mean and maximum lengths of 274 and 2072

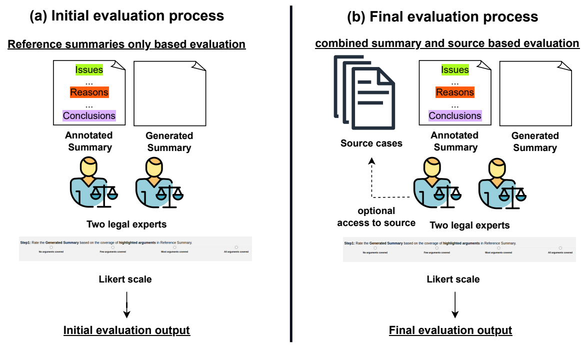

Figure 1: **Evaluation Process:** (a) *Initial evaluation* with human-annotated summaries and highlighted arguments. (b) *Final evaluation* with an option to cross-check the reference.

words, respectively. This subset has been extensively used in abstractive summarization research, particularly for constructing argument-aware abstractive summaries of legal opinions (Elaraby and Litman, 2022; Elaraby et al., 2023). The annotated argument roles follow the structure proposed in Xu et al. (2020, 2021), which breaks legal argument roles into three components: **Issue** (legal questions addressed by the court in the document), **Reason** (explanations for the court's decisions), and **Conclusion** (the court's rulings on the issues). Although these argument components constitute a small portion of the source cases, they typically account for $\approx 60\%$ of the summaries on average (Elaraby et al., 2023), highlighting the significance of considering argument roles in summary generation.

We considered the output of three different abstractive models in our evaluation process: (1) **Finetuned LED-base:** This model serves as the baseline for legal opinion summarization, as described in Elaraby and Litman (2022). It finetunes the pretrained longformer-encoder-decoder (Beltagy et al., 2020) on the CanLII cases without additional information about the argument structure of the document. (2) **arg-LED-base:** Utilizing the longformer encoder-decoder architecture, this model highlights argument units (Issues, Reasons, and Conclusions) with special tokens during both training and inference, as detailed in Elaraby and Litman (2022). (3) **arg-aug-LED-base:** This model extends the arg-LED-base model, as discussed in Elaraby et al. (2023). It incorporates a mechanism for sampling summaries during inference and selecting the best model that exhibits the highest overlap with the input case's predicted argument roles.

[1]Data obtained through an agreement with CanLII (`https://www.canlii.org/en/`).

4. Argument Coverage Evaluation

We relied on two legal experts (two co-authors who are lawyers) to perform our human evaluation process, which was conducted in two phases. *Figure 1 shows an overview over the initial evaluation process (a) and the final evaluation process (b).*

4.1. Initial Evaluation Process

Initially, as shown in Figure 1 (a), we chose not to provide the full legal opinion due to its lengthy nature and the sparse distribution of argument roles across the case. Instead, experts were provided solely with human-written summaries, predominantly comprising argument roles. We highlighted the types of argument roles within the summaries to aid evaluators in distinguishing between argumentative and non-argumentative sections.

Our evaluation guidelines incorporate a 4-point Likert scale, facilitating a detailed assessment of argument coverage within the summaries. A rating of 4 indicates a perfect coverage of argument components, while a rating of 1 denotes a complete absence of coverage. To minimize misinterpretation of each score, we provided definitions for each rating category. During this phase, we utilized human-annotated summaries from 5 distinct legal opinions randomly selected from CanLII cases. For each case, we sampled summaries from the three distinct LED models, resulting in a total of 15 cases and summary pairs. Upon completion by both experts, the weighted quadratic kappa agreement, calculated using the sklearn implementation [2], between the two experts reached 0.466.

Discrepancies between the two experts were examined in a separate session, revealing that most disagreements stemmed from confusion regarding whether a certain argument within the generated summary was stated differently in the source document.

4.2. The Final Evaluation Process

To address evaluators' disagreements in the initial evaluation phase, we provided evaluators with human-written summaries, as outlined in the initial process. Additionally, evaluators were given the option to cross-check whether a specific argument was stated differently in the source document, as illustrated in Figure 1 (b).

Legal expert evaluators were provided with 15 additional summaries drawn from 5 new legal opinions. Our evaluation results suggest that by offering this option alongside the human-written summaries, the overall weighted quadratic kappa agreement improved to 0.607. *The final evaluation guidelines are presented in Appendix A.*

4.3. Streamlining the Evaluation Process with Dedicated Software

To facilitate the experts' task, we developed a dedicated software for the longform evaluation of generated summaries. Our software builds upon the base code of the *Falte* tool (Goyal et al., 2022a), with several key enhancements: (1) **Keeping Expert State:** Recognizing the need for multiple sessions, we maintain the evaluation status for each expert, allowing them to complete the task across several sessions at their convenience. (2) **Inclusion of Likert Scale:** We include Likert scale definitions for each evaluation sample, aiming to reduce rating variability. (3) **Source Accessibility:** Acknowledging the positive impact of including source documents on the evaluation agreement, we added an option for experts to navigate to the source document. This allows them to cross-check confusing points against the source, improving accuracy. (4) **Highlighting Argument Roles:** To streamline the evaluation process, we highlight annotated argument roles in both the reference summaries and the source document. This facilitates cross-checking the generated summaries against them, reducing confusion. This approach is akin to solutions proposed by Krishna et al. (2023); Liu et al. (2023); Min et al. (2023), where evaluators are provided with atomic units of the summaries for evaluation. In our work, argument roles serve as the salient atomic units. *The tool is deployed and available online[3], enabling experts to complete tasks asynchronously. A screenshot is included in Appendix B [4].*

5. Results and Analysis

The final evaluation set consisted of 90 distinct generated summaries, that weren't included in the training phase, evenly selected from the three LED-based models, covering 30 unique legal opinion cases. Ratings were collected over two weeks using our dedicated software.

5.1. Experts' Agreement

The final quadratic kappa agreement was 0.483, which was lower than that obtained during the evaluation of the final evaluation process. *We hypothesize that this decline may be attributed to novel issues arising that were not addressed during the*

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

[3] https://summary-evaluation.herokuapp.com/

[4] https://github.com/EngSalem/legal-falte

Metrics	τ correlation coeff.		
	Expert 1	Expert 2	Average
rouge-1	0.35	0.33	0.37
rouge-2	0.33	0.30	0.33
rouge-L	0.28	0.34	0.34
BERTscore	0.31	0.29	0.33

Table 1: Automatic metrics correlations in *kendal tau* τ with legal expert evaluations. All τ values are statistically significant with $p < 0.01$.

training phase but required attention in the human guidelines. We also evaluate the agreement between expert rankings of summaries by computing Kendall's tau (τ) correlation coefficients. The final τ correlation coefficient is 0.429 with $p < 0.001$, indicating a significant pairwise agreement between ratings of different systems.

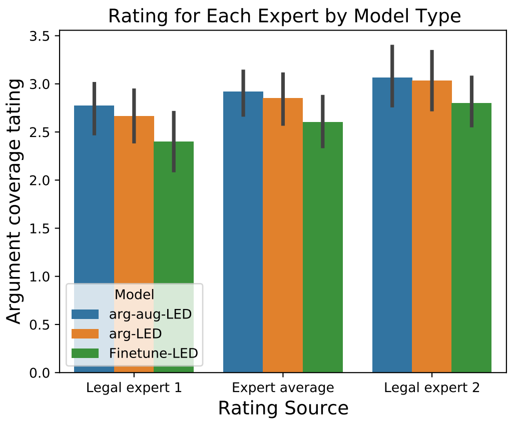

Figure 2: Average ratings. *Expert average:* average of Legal expert 1 and Legal expert 2.

5.2. Argument Aware Model Rankings

We analyzed the average rankings of summaries generated by different LED models. Figure 2 illustrates that the **Finetune-LED** model consistently received lower rankings from both legal experts compared to the **arg-LED** model (Elaraby and Litman, 2022), which highlights argument roles with special tokens, and the **arg-aug-LED** model (Elaraby et al., 2023), which leverages second-stage reranking to select the model with the highest argument similarity to the input. These findings are consistent with the significant correlation of rankings between both models discussed in 5.1, indicating that despite the drop in kappa agreement, experts agreed on the average rankings of summaries generated by different systems. *These results highlight that considering the argumentative components in the input document improves argument coverage in the generated summaries.*

5.3. Correlation with Automatic Metrics

We assess the effectiveness of automatic metrics previously employed in evaluating legal opinion summarization (Elaraby et al., 2023; Elaraby and Litman, 2022; Zhong and Litman, 2023) against human evaluation scores of argument coverage. These models primarily utilized ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) to assess their proposed approaches. Table 1 shows that ROUGE demonstrated relatively higher correlation scores, ranging from 0.34 to 0.37, compared to BERTScore. Nevertheless, these findings suggest the potential for developing metrics specifically tailored to capture argument coverage. For instance, Fabbri et al. (2021) showed stronger correlations with aspects like fluency, consistency, coherency, and relevance, underscoring the need for more targeted metrics for assessing argument coverage.

5.4. Abstractiveness and Length of Summaries Effect on Ratings

Abstractiveness was quantified by computing the percentage of novel n-grams in each summary (See et al., 2017). Our findings, presented in Table 2, indicate that overall abstractiveness has limited influence on the ratings. However, as the number of novel n-grams increases (case of 4-gram), it can have a negative impact on argument coverage.

Novel n-grams	Average	Expert 1	Expert 2
1-gram	-0.182*	-0.151	-0.180
2-gram	0.002	0.001	0.001
3-gram	-0.045	-0.095	0.002
4-gram	-0.200*	-0.251*	-0.129

Table 2: τ values for novel n-grams vs ratings. * refers to $p < 0.05$.

Given the variability in our summary lengths, we aim to investigate its influence on argument coverage ratings. However, Table 3 indicates that the length of the summary has no significant effect on argument coverage.

Expert Average	Expert 1	Expert 2
0.01	0.12	-0.08

Table 3: τ values for summary length vs ratings. All values are with $p > 0.05$.

6. Conclusion

In this paper, we explored the concept of *argument coverage*, a new aspect in the evaluation of abstractive summarization. Our focus was primarily on long legal opinions, where ensuring thorough argument coverage is essential for producing meaningful summaries. We introduced specific evaluation guidelines crafted for assessing argument coverage, allowing us to re-evaluate existing models for long legal opinion summarization. Our findings underscored the efficacy of integrating argument roles into the summarization process. Furthermore, we examined the automatic summarization metrics commonly used in legal opinion summarization research. Although ROUGE emerged as the most promising metric, our analysis suggests the potential for developing dedicated automatic metrics tailored to assess argument coverage more effectively. In future research, we aim to incorporate argument role types for a more nuanced evaluation and explore more efficient automatic metrics.

Limitations

One limitation of this study is the absence of exploration into generated summaries from Large Language Models, which represents a promising avenue for future research in legal opinion summarization. Additionally, a larger dataset of legal opinions could have been incorporated into the evaluation training to refine the evaluation guidelines and potentially mitigate disagreements between experts more effectively. This would enhance the robustness of the evaluation process and bolster the reliability of the results. Moreover, while the focus was on legal opinions, extending the evaluation study to other domains where argument coverage is crucial, such as debates, would provide more comprehensive and inclusive guidelines for summarization.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2040490 and by Amazon. We would like to thank the members of both the Pitt AI Fairness and Law Project and the Pitt PETAL group, as well as the anonymous reviewers, for valuable comments in improving this work. We would also like to thank Ahmed Ismail Zahran for his feedback on developing the annotation tool.

7. Bibliographical References

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

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, 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.

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 R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022a. FALTE: A toolkit for fine-grained annotation for long text evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–358, Abu Dhabi, UAE. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022b. SNaC: Coherence error detection for narrative summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 444–463, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Wojciech Kryscinski, 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.

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.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Ankan Mullick, Abhilash Nandy, Manav Kapadnis, Sohan Patnaik, Raghav R, and Roshni Kar. 2022. An evaluation framework for legal document summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4747–4753, Marseille, France. European Language Resources Association.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Olivier Salaün, Aurore Troussel, Sylvain Longhais, Hannes Westermann, Philippe Langlais, and Karim Benyekhlef. 2022. Conditional abstractive summarization of court decisions for laymen and insights from human evaluation. In *Legal Knowledge and Information Systems*, pages 123–132. IOS Press.

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.

Huihui Xu and Kevin Ashley. 2023. A question-answering approach to evaluating legal summaries. In *Legal Knowledge and Information Systems*, pages 293–298. IOS Press.

Huihui Xu, Jaromír Šavelka, and Kevin D Ashley. 2020. Using argument mining for legal text summarization. In *Legal Knowledge and Information Systems*, pages 184–193. IOS Press.

Huihui Xu, Jaromir Savelka, and Kevin D Ashley. 2021. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 250–254.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yang Zhong and Diane Litman. 2023. STRONG – structure controllable legal opinion summary generation. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 431–448, Nusa Dua, Bali. Association for Computational Linguistics.

A. Final Evaluation Guidelines

Table 4 shows the final evaluation guidelines provided to legal experts to obtain argument coverage ratings.

B. Evaluation Tool

Figure 3 shows a snippet from the evaluation tool used for collecting argumentation coverage.

Guide for Evaluation: Argument Coverage
Description
Argument Coverage: Do generated summaries cover the important points of the reference summary? You will be asked to rate the generated summary on a 4-point Likert scale to assess how well it covers the arguments in comparison to the highlighted arguments in the reference summary, which represent ground truth.
Recommended Steps
• Spend time to first read the reference summary until you understand the highlighted arguments. • Read the generated summary until you understand its contents. • Identify whether each argument highlighted in the reference summary is covered in the generated summary. • If in doubt about a certain argument in the generated summary, click on the "go to source" button to double-check it against the source.
Rating scale of the Generated Summary
1. **No arguments covered:** The generated summary did not cover the highlighted arguments in the reference summary or covered them only inadequately. 2. **Few arguments covered:** The generated summary adequately covered only a limited number of the highlighted arguments in the reference summary. 3. **Most arguments covered:** The generated summary adequately covered most of the arguments highlighted in the reference summary. 4. **All arguments covered:** The generated summary adequately covered all the highlighted arguments in the reference summary.

Table 4: Final evaluation guidelines for argument coverage.

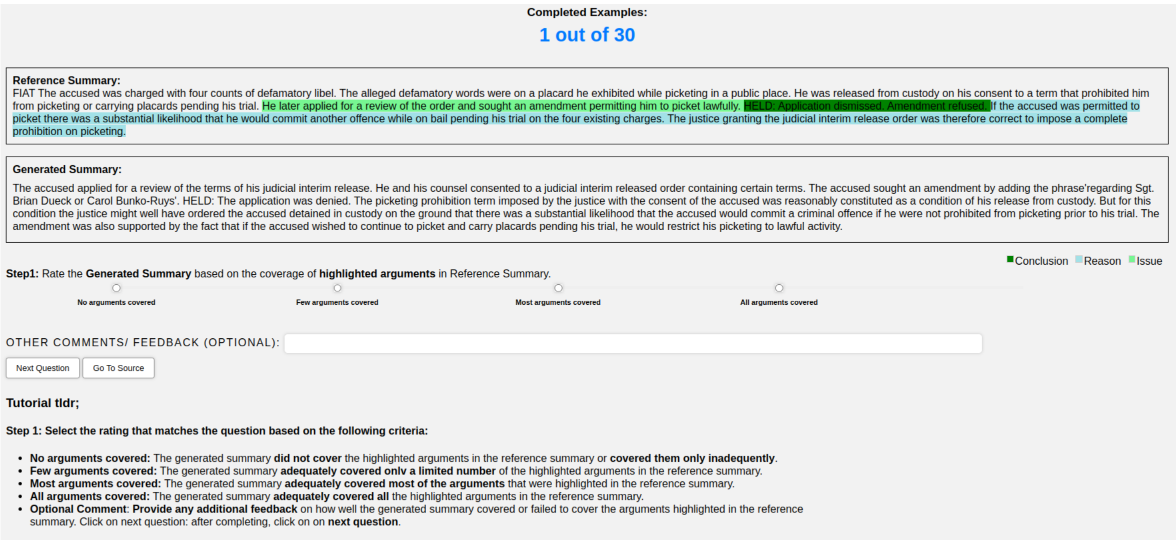

Figure 3: Screenshot from the tool used to collect argument coverage ratings from experts.