CIVILSUM: A Dataset for Abstractive Summarization of Indian Court Decisions

Anonymous ACL submission

Abstract

We introduce CIVILSUM, a dataset of 23,350 legal case decisions paired with human-written abstractive summaries from the Supreme Court of India and Indian High Courts. In contrast to other domains such as news articles, our analysis shows the most important content tends to appear at the end of the documents. We measure the effect of this *tail bias* on summarization performance using strong baselines for long-document abstractive summarization, and the results highlight the importance of long sequence modeling for the proposed task. CIVILSUM and related code are publicly available for research purposes.[1]

1 Introduction

With the growing demand for automation of legal systems, the development of natural language processing (NLP) techniques for analyzing legal documents has become a critical area of research (Dale, 2019; Chalkidis et al., 2020; Zhong et al., 2020; Moreno-Schneider et al., 2020). In particular, summarizing legal documents is an important and challenging problem due to their length and technical complexity. These characteristics increase the difficulty and cost for the collection of high-quality reference summaries required by state-of-the-art supervised summarization approaches.

To address these challenges, we introduce CIVILSUM, a dataset for abstractive summarization of legal documents. CIVILSUM comprises a collection of 23,350 legal case decisions from the Supreme Court of India and other Indian High Courts, each paired with a summary written by a legal professional. The dataset provides a rich source of information for training and evaluating NLP models for legal summarization tasks.

In this work, we describe the process of constructing the CIVILSUM dataset and compare its quantitative and qualitative characteristics with previous work in this domain. Our analysis reveals an interesting observation that the most important summarizable content tends to appear at the end of the legal documents, which is opposite from the lead bias observed in other domains such as news articles (Nallapati et al., 2016; Narayan et al., 2018). We also evaluate our dataset using two architectures for long-document abstractive summarization, namely Longformer (Beltagy et al., 2020) and FactorSum (Fonseca et al., 2022). Our results reveal that abstractive approaches outperform paragraph-based extractive methods, emphasizing the need for fine-grained, intra-paragraph abstractive processing to generate high-quality summaries on the CIVILSUM dataset. Our findings also suggest that the end of documents contains more informative content, as observed by comparing the results with lead and tail content guidance. Given recent advances in

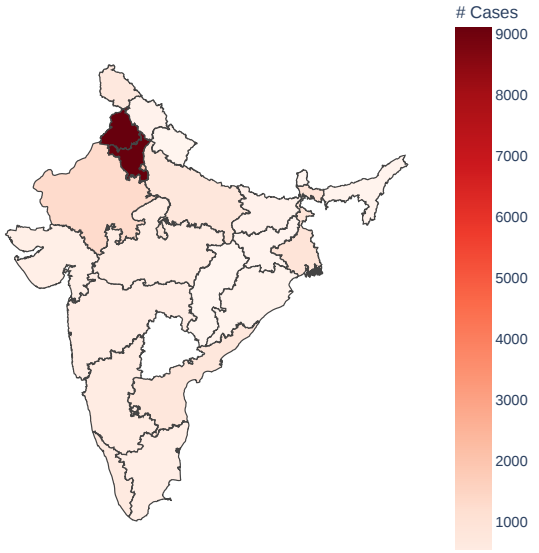

Figure 1: Distribution of CIVILSUM legal cases across Indian states. The majority of samples originate from the Supreme Court of India (4,499; not on the map) and the High Courts of Pubjab and Haryana (9,111), and Delhi (1,790).

[1]Link removed to preserve anonymity. We release our corpus under the CC BY-NC-SA 4.0 license.

Dataset	# docs (train/val/test)	Document		Summary		% novel n-grams in summary			
		words	sents	words	sents	1-gram	2-gram	3-gram	4-gram
IN-Abs (Shukla et al., 2022)	7,030/-/100	4,378	-	1,051	-	18.95	34.71	47.19	56.12
EUR-LexSum (Klaus et al., 2022)	3,447/689/459	11,864	340	1,011	32	43.70	71.00	84.62	90.29
CivilSum	21,015/1,168/1,167	2,123	90	104	4.5	62.60	91.52	98.87	99.77

Table 1: Statistics for legal summarization datasets, including number of documents, average length in words/sentences and summary abstractiveness (measured as percentage of novel n-grams).

large language models (LLMs) and their effectiveness in news summarization (Zhang et al., 2023), we also assess our dataset using Llama 2 (Touvron et al., 2023), an open-source LLM with the capacity to model lengthy text. Although ROUGE performance is inferior to the other two methods, human evaluation of overall summary quality shows Llama 2 summaries were more favored.

2 Related Work

While most of the legal NLP work focuses on US datasets, other jurisdictions are also studied, including summarization of 4,595 curated European regulatory documents (EUR-LexSum; Klaus et al., 2022), and topic modeling applied to multi-document summarization in the Brazilian lawmaking process (Silva et al., 2021). An argument mining approach is used to improve abstractive summarization of Canadian legal cases (Elaraby and Litman, 2022). Their proposed dataset consists of 1,262 legal cases obtained through an agreement with the Canadian Legal Information Institute[2].

In the Indian context specifically, similar summarization problems have been explored (Bhattacharya et al., 2021; Shukla et al., 2022; Ghosh et al., 2022). Bhattacharya et al. (2021) provided a dataset and performed extractive summarization operations on the dataset. Shukla et al. (2022) developed three datasets, primarily "IN-Abs" consisting of 7,130 document-summary pairs obtained from the website of the Legal Information Institute of India[3], "IN-Ext" consisting of 50 manually annotated summaries of judgments, and "UK-Abs" from the website of UK Supreme court[4] having 693 cases. For those datasets, they perform and evaluate both extractive and abstractive summarization models. Our work aims to increase the scale of summarization datasets in the Indian legal domain.

[2] https://www.canlii.org/en/

[3] http://www.liiofindia.org/in/cases/cen/INSC/

[4] https://www.supremecourt.uk/decided-cases/

3 Dataset Construction

The focus of the CivilSum dataset is on civil cases heard by the Supreme Court of India and Indian High Courts from the country's independence (1947) up until the 2010–2011 calendar year. In comparison to previous work, CivilSum is significantly larger in dataset size. In addition, our human-written abstractive summaries have a higher compression ratio, providing more concise and informative summaries. The compression ratio is calculated as the ratio of the number of words in the original document to the number of words in the summary. Previous datasets have a compression ratio of around 5-10. In contrast, the summaries in CivilSum have a higher compression ratio of around 16, making the task of summarization more challenging. Following Narayan et al. (2018), we also compute the fraction of n-grams in the summary that are not present in the original document. A summary of the main dataset statistics compared to existing legal summarization datasets is provided in Table 1. The distribution of cases per state is illustrated in Figure 1.[5]

3.1 Data Preparation

We provide the details of our data collection and cleaning steps in Appendix A.

3.2 Paragraph Reference Extraction

A salient stylistic feature of the dataset is that most of the paragraphs in the summaries include textual references to the relevant paragraphs in the judgments, which we hypothesize is an important signal for summarization modeling. To leverage this data, we devised a pattern-matching algorithm to extract paragraph references of the form `[Paras 17, 10, 15]` (refer to Appendix C for an example). By applying this heuristic, we create a dataset

[5] This map was generated using the Plotly tool and may not include disputed regions or other areas of contention. The boundaries and names shown on the map do not imply official endorsement or acceptance by the authors or their affiliations.

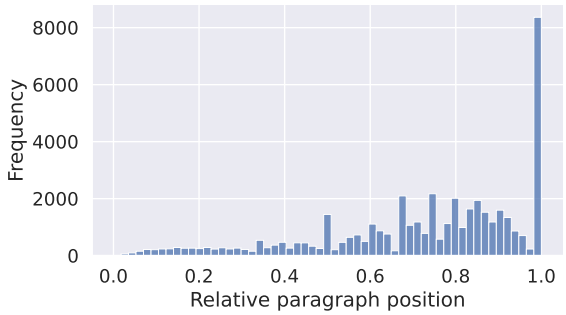

Figure 2: Distribution of relevant paragraph positions in the documents (training split) exhibiting tail bias.

where each paragraph in a judgment is labeled as 1 if mentioned in the summary, and 0 otherwise. Out of 23,350 documents in the dataset, 22,682 ($\approx 97\%$) contain at least one referenced paragraph in the reference summaries.

This paragraph reference information reveals an interesting insight about the information distribution in the dataset: most of the relevant content is located towards the end of the documents, a characteristic we refer as to *tail bias* (Figure 2). A consequence of this finding is that summarization systems that are biased towards leading information, as commonly seen in news summarization (Grenander et al., 2019; Zhu et al., 2021), should not perform well on our benchmark. We explore this tail bias in various settings in Section 4.

4 Methodology and Experiments

We describe summarization experiments with various types of architectures designed to process long documents. Our objective is to provide a baseline performance assessment for future work and to measure how the distribution of relevant information in the documents affects summarization performance. The models are detailed as follows:

Random extractive baseline. To get an estimate for the task difficulty, we randomly sample paragraphs from the documents up to 7% of the total words, subject to a minimum of 150 words. If the document has 150 or fewer words, the entire document is used as the summary.

Extractive oracle paragraphs. We also obtain oracle extractive summaries that include only paragraphs mentioned in the reference summaries (refer to Section 3.2 for details). The budget constraints are the same as the random extractive baseline described above.

FactorSum (Fonseca et al., 2022), an abstractive summarization model that employs a sampling mechanism to generate several summary snippets (*summary views)*, which are then combined into a final summary following a guidance optimization objective. We leverage guidance to bias the resulting summary to focus on the start of the document (*lead guidance*) and on the end of the document (*tail guidance*). Additionally, we measure the performance using extractive oracle paragraphs as guidance, that is, we encourage the final summary to be similar (using ROUGE-1 as the similarity metric; see below) to oracle paragraphs. We choose FactorSum because it can handle long documents by relying on a relatively small sequence-to-sequence backbone (BART-base; Lewis et al., 2020) and a short input context (1,024 tokens). See Appendix B for additional details.

Longformer (Beltagy et al., 2020), a transformer-based model that implements an attention mechanism that scales linearly with the input length, which makes it suited for the processing of the long documents from our dataset. We experiment with various input configurations, including 4,096 input tokens and truncated documents of both the first and last 1,024 tokens. Additionally, we test the performance of the model when using only oracle paragraphs as inputs. See Appendix B for additional details.

Llama 2 (Touvron et al., 2023), a transformer-based large language model. With up to 70 billion parameters, Llama 2 has the capacity to process lengthy texts. We leverage the finetuned chat version of Llama 2 and provide it with 4,096 tokens as input. See Appendix B for additional details.

5 Results

Automatic Evaluation We measure performance by ROUGE-1/2/L F1 score (Lin, 2004), following previous work in the summarization literature. These metrics measure the word overlap, bigram overlap, and longest common sequence between system-generated and reference summaries. The results in Table 2 show a large gap in performance between a paragraph-based extractive summarizer and the abstractive approaches. This result suggests that summarizing more fine-grained, intra-paragraph abstractive processing is required to generate high-quality summaries. Still, we can verify that paragraph references are highly informative,

Model	Input Tokens	R-1	R-2	R-L
Extractive (random)	-	31.72	9.02	21.38
Extractive (paragraphs)	-	32.75	10.53	22.29
FactorSum (lead)	1024	40.33	15.74	31.98
FactorSum (tail)	1024	41.80	16.53	33.30
FactorSum (paragraphs)	1024	**46.51**	**20.67**	**37.07**
Longformer	4096	44.80	18.37	36.85
Longformer (lead)	1024	41.89	15.97	34.45
Longformer (tail)	1024	43.80	17.37	35.85
Longformer (paragraphs)	1024	42.25	15.78	34.51
Llama-2-chat-7B	4096	37.12	12.55	25.43
Llama-2-chat-13B	4096	36.73	11.63	25.61
Llama-2-chat-70B	4096	37.39	12.61	25.74

Table 2: ROUGE F-1 scores for the summarization task. *lead* and *tail* refer to summaries focusing on the start and end of documents respectively. The *paragraphs* leverage information from oracle paragraphs as described in Section 4.

improving the scores in ≈ 14 R-1 over the random extractive summarizer, and ≈ 6 R-1 over FactorSum with lead guidance.

Another salient pattern in the results is the higher informativeness towards the end of the documents, which can be verified by comparing the results of FactorSum with lead and tail guidance. Similarly, we observe a strong loss in Longformer performance by truncating the documents to the first 1,024 tokens (lead) compared to using 4,096 tokens, but the loss in performance is much smaller when using the last 1,024 tokens (tail). Finally, we observe that Llama 2-chat exhibits superior performance to extractive summarization approaches, yet remains inferior to other abstractive methods. We posit this stems from evaluating Llama 2 in a zero-shot setting without fine-tuning. As LLMs show promise on legal summarization, we leave finetuning Llama 2 with in-domain data to future work. Additionally, we observe that scaling Llama 2-chat parameters does not further improve performance. Nonetheless, the zero-shot results demonstrate Llama 2-chat's capability to generate reasonable abstractive summaries without training. Further tuning could likely adapt the model to the target summaries' style and content.

Human Evaluation In addition to automated measures like ROUGE, we designed a human evaluation to collect preference annotations. For each given document, annotators were presented summaries from all three summarization systems (FactorSum, Longformer and Llama 2). They were first instructed to select their most preferred summary to replace a technical judgement abstract. Subsequently, they were prompted to choose the best summary accounting for criteria such as informativeness and fluency. Refer to Appendix D.1 for further details. Our evaluators comprised two trained Indian lawyers familiar with the cases, who examined 25 randomly selected samples.

Regarding the results for the first question, the first annotator preferred the FactorSum, Longformer, and Llama 2 summaries 10, 9, and 6 times, respectively. The second annotator preferred them 8, 13, and 4 times. The inter-annotator agreement as measured by Cohen's kappa was 0.44, indicating moderate agreement. For the second question, the preferences were 6, 6, 13 and 6, 7, 12, respectively. The inter-annotator agreement was 0.52, again suggesting moderate alignment. These results imply that for technical adequacy, summaries from supervised models like FactorSum and Longformer were preferred. However, considering overall summary quality, the Llama 2 summaries were favored.

In addition, the annotators observed that although the summaries were generally adequate and captured key points successfully, there were deficiencies in sentence construction ambiguity, erroneous interpretations of interest payment, and sporadic incompleteness. Concerns were also raised about conciseness, omission of conclusions, overuse of constitutional articles, indirect addressing of the real issue, use of personal pronouns, and insufficient consideration to legal aspects. We provided a detailed discussion in Appendix D.2. Overall, the evaluation suggests that our dataset is challenging and current summarization systems struggle to produce satisfactory summaries.

6 Conclusions and Future Work

In this work, we introduce CIVILSUM, a novel dataset for legal summarization containing 23,350 court decisions paired with human-written summaries. We describe the steps for dataset construction and provide extractive and abstractive summarization baselines to serve as a benchmark for further investigation. We explore stylistic features of the documents such as paragraph references and measure how information tail bias affects the summarization performance in diverse settings. A promising direction for future work would be to assess the factuality of generated summaries in terms of relevant entities such as legislation references, which is a crucial aspect of court decisions.

Limitations

One limitation of our study is that we did not explicitly address the issue of hallucinations introduced by the abstractive summarization models. Hallucinations refer to the generation of inaccurate or misleading information in the generated summaries. While our dataset, CIVILSUM, presents a challenging task for abstractive summarization, the presence of hallucinations in the model outputs (mostly in the form of non-factual references to paragraphs and legal articles) indicates the need for additional research and development to improve the reliability and trustworthiness of the summarization process.

In addition, we did not fully explore or utilize information like presences and mentions of legal acts and references to previous judgments. These aspects of legal documents often contain informative content that could contribute to the creation of more accurate and comprehensive summaries. By incorporating such information, future research could potentially enhance the summarization process and generate summaries that better capture the legal context and implications.

Another limitation arises from the computational budget constraints we faced. Due to these constraints, we focused our evaluation on two state-of-the-art architectures for long-document abstractive summarization: Longformer and FactorSum. While these models demonstrated promising performance on the CIVILSUM dataset, we acknowledge that other models, such as Long-T5, were not included in our evaluation. Furthermore, for LLMs, despite evaluating our dataset with Llama 2, the appraisal was undertaken in a zero-shot manner without fine-tuning. Further investigation into the performance of these alternative models, including further fine-tuning of LLMs with in-domain data, could provide insights and potentially lead to even more effective summarization approaches for legal documents.

Ethics Statement

The development and application of NLP techniques for analyzing legal documents, as described in this work, raise important ethical considerations. As researchers, we recognize the need to address these ethical implications and ensure that our work adheres to the principles of responsible research and practice. In this ethics statement, we outline our approach to ethical considerations and discuss the potential impact of our work.

Data Collection and Usage The CIVILSUM dataset, introduced in this paper, comprises legal case decisions from the Supreme Court of India and other Indian High Courts, along with summaries written by legal professionals. It is essential to highlight that the collection and usage of legal documents do not involve user-related or private data as the legal case decisions are publicly available. For more details about the data and copyright, please refer to Appendix A.1. We want to emphasize that our data collection process adheres to appropriate legal and ethical guidelines. We are committed to ensuring that the dataset is used solely for research purposes and that any potential biases or discriminatory elements are minimized. We release the dataset under the CC BY-NC-SA 4.0 license[6].

Bias and Fairness Given the nature of legal documents and their potential impact on individuals and society, it is crucial to address biases and promote fairness in the development and evaluation of NLP models for legal summarization. We acknowledge that biases can be inherent in the legal system and may be reflected in the dataset itself. We encourage researchers and practitioners to be vigilant in their analysis, interpretation, and application of the CIVILSUM dataset to ensure fairness and equity.

Human and Legal Considerations Legal documents often involve sensitive information about individuals and legal matters. It is imperative to respect the confidentiality and privacy of the parties involved. Our study focuses on the analysis and summarization of publicly available legal case decisions while ensuring that personal information is appropriately protected. We urge researchers and practitioners working with legal documents to adhere to relevant legal and ethical guidelines and consult with legal professionals to ensure compliance with data protection laws and regulations specific to their jurisdiction.

References

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

Paheli Bhattacharya, Soham Poddar, Koustav Rudra, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. Incorporating domain knowledge for extractive summarization of legal case documents. In *Proceedings of*

[6] https://creativecommons.org/licenses/by-nc-sa/4.0/

the eighteenth international conference on artificial intelligence and law, pages 22–31.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Robert Dale. 2019. Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25(1):211–217.

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.

Marcio Fonseca, Yftah Ziser, and Shay B. Cohen. 2022. Factorizing content and budget decisions in abstractive summarization of long documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6341–6364, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Satyajit Ghosh, Mousumi Dutta, and Tanaya Das. 2022. Indian legal text summarization: A text normalisation-based approach. *arXiv preprint arXiv:2206.06238*.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. 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 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Svea Klaus, Ria Van Hecke, Kaweh Djafari Naini, Ismail Sengor Altingovde, Juan Bernabé-Moreno, and Enrique Herrera-Viedma. 2022. Summarizing legal regulatory documents using transformers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2426–2430.

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.

Julian Moreno-Schneider, Georg Rehm, Elena Montiel-Ponsoda, Víctor Rodriguez-Doncel, Artem Revenko, Sotirios Karampatakis, Maria Khvalchik, Christian Sageder, Jorge Gracia, and Filippo Maganza. 2020. Orchestrating NLP services for the legal domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2332–2340, Marseille, France. European Language Resources Association.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

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.

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.

Nádia FF da Silva, Marília Costa R Silva, Fabíola SF Pereira, João Pedro M Tarrega, João Vitor P Beinotti, Márcio Fonseca, Francisco Edmundo de Andrade, and André CP de LF de Carvalho. 2021. Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 104–120. Springer.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan

Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1462–1471.

A Data Preparation

A.1 Copyright Information

In accordance with the provisions of the Indian Copyright Act, 1957, it is affirmed that the judicial pronouncements are readily accessible and can be accessed through the website[7] by conducting a search using the name of the specific case. It should be noted that the headnotes or summaries of these judicial pronouncements are protected under the Indian Copyright Act, 1957, with copyright belonging to Copyright © 2016 Patiala Law House[8].

Furthermore, this dataset's license is restricted to specific purposes such as conducting academic or educational research or study. It should be duly acknowledged that the utilization of the judicial pronouncements from the aforementioned website is carried out within the confines of the license provided, and thus does not infringe upon the provisions set forth by the copyright act.

A.2 Data Collection

The legal cases were obtained from the Patiala Law House, Patiala, India, through an agreement. The data consists of judgments from various judicial courts situated in different parts of India immediately after the independence of India till the calendar year 2010-11. For each case judgment, we obtain the following information (in DOCX format): 1) A document identifier derived from the position on the hit list returned by the system; 2) The petitioner and respondent's and contesting parties' names. It is important to notice that the 'versus' clause may contain numerous petitioners' and respondents' names; 3) The name of the court that rendered this ruling; 4) The judgment's summary, usually referred to as a headnote; 5) References to previous relevant cases. These references pertain to the legal cases cited by the adjudicating authority based on earlier judgments referred to by the parties in the case; 6) The text conveying the judge's decision; 7) A summary of the judgment. The summaries are condensed from the judgments and are manually written, with no automation involved.

A.3 Data Cleaning

The data is preprocessed and cleaned, starting from the DOCX files, which are the original formats of the judgments. We developed a matching algorithm to recognize six different types of informa-

[7] https://indiankanoon.org/

[8] https://patialalawhouse.blogspot.com/

tion: ranks, names of the contesting parties, name of the court where the judgment appeared, references from previous judgments (from which the current judgment draws support for its claims), judgment text, and summaries, which are headnotes in legalese. The written documents contain errors, thus several edge cases are addressed. For instance, documents are divided by words with spelling variations like `ORDER`, `COMMON ORDER`, `JUDGMENT`, `JUEDGMENT`, `JUDGEMENT`, and `JUDGMEN2T` to separate judgment content from the rest of the text. The names of the contesting parties receive a similar level of attention, and the strings `cases referred` and `case referred` are used to separate reference cases in the judgment document because this is a common pattern found after carefully assessing a subset of decisions. Our matcher misses certain faults because the documents were prepared manually by human experts and are therefore prone to human error.

B Additional Details for Experiments

For FactorSum, we augment the document-summary pairs by creating pairs of *document views* and *summary views* that capture different perspectives of the original documents. To this end, we first perform sentence tokenization on both documents and summaries. Then, we uniformly sample 20% of the sentences in the documents to serve as document views for each one of the 21,013 documents in the training set, resulting in 420,260 shorter training samples. Each document view is paired with a corresponding subset of the original summary, which we refer to as a summary view. Using the same approach, we obtain 23,360 and 23,340 document-summary view pairs for the validation and test sets respectively. Apart from the usual input truncation in transformer models, no further preprocessing is performed for Longformer and Llama 2.

We use a BART-base (Lewis et al., 2020) checkpoint from HuggingFace[9] as starting point to train FactorSum summary views generator. The maximum length for generation per view is set to 128 tokens, the effective batch size is 64, and we use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 5×10^{-5}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The training is performed for 50,000 steps on 4 GeForce GTX 1080 Ti GPUs, and we choose the checkpoint with the highest ROUGE-1 F1 score on the validation split. We employ a pre-trained LED-base checkpoint from HuggingFace[10] for Longformer and finetuned the model using a learning rate of 1×10^{-4} on 4 NVIDIA A100 GPUs with 128 effective batch size. The maximum length for summary is set to 256 tokens. All other training details are the same as those used for FactorSum. During inference, FactorSum performs the greedy optimization described by Fonseca et al. (2022) using the same sampling hyperparameters as the training phase (20 document views per sample, each with 20% of the original sentences), with a budget constraint of 190 words per summary. Longformer uses a beam size of 3. For Llama 2-chat, we query the model with the the prompt template: "`${document}.\n Write a summary of the text above in 4 sentences.`", and parse the model's completion as the candidate summary. For sampling hyperparameters, we use a value of 0.6 for temperature, and 0.9 for top-p filtering.

C Data and Summary Samples

We provide samples of our dataset in Table 3 and Table 4. Each summary paragraph starts with a supporting legislation and usually ends with references to relevant paragraphs from the source documents (shown in blue color). We also provide samples of generated summary from Longformer and FactorSum in Table 5 and Table 6.

D Additional Details for Human Evaluation

D.1 Human Evaluation Guidelines

In order to assess the quality of summaries written for legal judgments, we conducted a human evaluation study. We use system generated summaries from FactorSum (with paragraph guidance), Longformer (with 4096 input tokens), and Llama 2-chat (70B) for the study. The purpose of this study was to gather subjective assessments from human evaluators based on specific guidelines. The guidelines were designed to evaluate the relevance, consistency, fluency, and coherence of the output summaries.

We asked human evaluators to find the answers to two questions for each summary pair which included summaries generated by FactorSum, Longformer, and Llama 2: Initially, we tasked human evaluators with selecting the superior summary

[9] https://huggingface.co/facebook/bart-base

[10] https://huggingface.co/allenai/led-base-16384

to replace a technical judgment abstract. Subsequently, the second question pertained to identifying the best summary overall, taking into account factors such as informativeness, fluency, and more. We defined the following evaluation criteria for human evaluators for the second question and instructed them to select the best summary based on these definitions:

1. **Relevance:** The rating measures how well the output summary captures the key points of the Judgment. Consider whether all and only the important aspects are contained in the output summary.

2. **Consistency:** The rating measures whether the facts in the output summary are consistent with the facts in the original Judgment. Consider whether the output summary reproduces all facts accurately and does not include untrue information.

3. **Fluency:** This rating measures the quality of individual sentences, whether they are well-written and grammatically correct. Consider the quality of individual sentences.

4. **Coherence:** The rating measures the quality of all sentences collectively, and how well they fit together and sound natural. Consider the quality of the output summary as a whole.

5. **Informativeness:** The rating measures whether the summary abstract encompasses all the essential details contained within the judgment.

We provided the following instructions to the human evaluators:

1. Carefully read the Judgment and be aware of the information it contains.

2. Read the three provided generated summaries.

3. Pick the best replacement for the reference legal summary.

4. Pick the best output summary on the five dimensions (Relevance, Consistency, Fluency, Coherence, Informativeness).

5. Consider the definitions provided for each criterion while rating the output summary.

D.2 Discussion of Human Evaluation Results

During the evaluation process, various strengths and weaknesses were identified in the generated summaries. Notable strengths included the overall acceptability of the summaries and their ability to effectively capture the key points of the judgment. While the recorded factual details were not entirely accurate, they were satisfactory overall. However, there were also identified weaknesses, such as instances of ambiguity in sentence construction, faulty interpretations regarding the payment of interest, and occasional incompleteness in the summaries.

Furthermore, concerns were raised about the lack of conciseness and occasional omission of conclusions, which are crucial elements in summarizing legal judgments. In the sample examples (see Table 5 and Table 6), the evaluator highlighted specific issues, including the overuse of articles 14 and 16 of the Constitution of India without proper contextual relevance, a tendency to refer to party names instead of directly addressing the real issue, and the use of personal pronouns instead of maintaining an objective tone. Moreover, there was a lack of sufficient attention to the legal aspects of the issue, resulting in an incomplete and inadequate portrayal of the real issue from a legal standpoint.

This study provides valuable insights into the effectiveness of summary writing for legal judgments, identifying specific strengths and weaknesses in the generated summaries. The findings emphasize the importance of clear and unambiguous sentence construction, accurate interpretation of information, completeness in summarizing key points, concise and straightforward language, inclusion of conclusions, proper contextual use of legal provisions, objective addressing of the real issue, and a comprehensive understanding of the legal aspects involved.

A. T.N. Recognised Private Schools (Regulation) Act, 1973, Sections 22, 23 and 24 - T.N. Recognised Private Schools (Regulation) Rules, 1974, Rule 17 - Termination - Approval - Enquiry conducted and conclusion made that proposed punishment was not warranted - No interference due to proper application of mind. **[Paras 12 & 13]**

B. T.N. Recognised Private Schools (Regulation) Act, 1973, Sections 22, 23 and 24 - T.N. Recognised Private Schools (Regulation) Rules, Rule 17 - Applicability - Termination - Approval - Approving authority must consider whether proved charges justify a particular action - Declined approval order does not suffer from any infirmity. **[Paras 12, 13]**

C. T.N. Recognised Private Schools (Regulation) Act, 1973, Sections 22, 23 and 24 - T.N. Recognised Private Schools (Regulation) Rules, Rule 17 - Termination - Validity - Backwages - Directed reinstatement with backwages not proper - Payment of 60% salary until superannuation date appropriate due to lack of other indicated aspects beyond section 22. **[Para 14]**

Arijit Pasayat, J. - Undaunted by reverses before the departmental authorities and the High Court, [...] The controversy lies within a narrow compass and factual position being undisputed, a brief reference thereto would suffice.

2. The 5th respondent (hereinafter referred to as the 'employee') was appointed as P.G. Assistant for teaching English in 1978. [...] On 29.8.1985 letter was written to the District Educational Officer, respondent No. 4 (in short the 'DEO') requesting for early action in the matter.

3. The DEO issued a notice to the employee but there was no response thereto. [...] In fact, the employee had not worked and abandoned work. But the DEO again directed the management to reinstate the employee and pay him back wages failing which the steps regarding direct payment were to be taken.

4. Aggrieved by these orders, the Management filed a writ petition before the Madras High Court. Learned Single Judge was of the view that in terms of what is required under Section 22(1) of the Act, [...] Another teacher has been appointed and the management is paying his salary.

⋮

11. The second plea of learned counsel for the management was even if the authority had jurisdiction, [...] It was strenuously contended that the welfare of the students' aspect was not even taken note of.

12. The role a teacher plays in shaping the career and future of a student needs no great emphasis. [...] This is because the approving authority has to consider whether the proved charges on the facts and the materials justify a particular action. Since reasons have been given on consideration of the materials, there is no scope for interference.

13. The order of the authorities declining to accord approval does not suffer from any infirmity. The High Court was justified in declining to interfere.

14. Another point urged by learned counsel for the appellant was that the direction for the back wages in its entirety is not justified because the employee absented from duty without sanctioned leave for long periods and even on some dates he went away during the school period and even abandoned the classes on several days. [...] No further orders are to be passed in the application for modification of earlier interim orders passed. The appeal is disposed of accordingly. Order accordingly.

Table 3: Sample abstract and Judgment from the CIVILSUM test set (ID = 648).

For Respondent No. 3. :- R.K. Malik, Advocate. A. Haryana Labour Department (Group A and Group B) Rules, 1987, Rules 9 and 7 - It is noted that the existing rules have been repealed and the Draft Service Rules framed and approved by Public Service Commission, but the draft rules have not been notified in Gazette and thus, cannot be considered as executive instructions. 1985(1) SLR 41, relied upon. **[Paras 7 and 8]**

B. Haryana Labour Department (Group A and Group B) Rules, 1987, Rules 7 and 9 - In relation to the constitutional validity of Article 16, seniority and acting promotion granted to the petitioner, it was established that the petitioner's promotion was regularised from 6.10.1986, but with no back salary. However, the respondent was appointed to the post with effect from 24.2.1984 and appointment regularised by Public Service Commission with effect from 11.1.1986, thereby proving that the respondent was senior to the petitioner. **[Paras 7 and 8]**

N.K. Kapoor, J. - The petitioner has sought issuance of writ of certiorari quashing promotion order Annexure P which Mange Ram stated to be junior to the petitioner has been promoted without considering the claim of the petitioner.

2. The petitioner joined the Labour Department as a Clerk in the year 1961 and after getting few promotions is presently working as Statistical Officer which is Class-II post. [...] Provided that their inter se seniority for purpose of consideration for promotion shall be on the basis of continuous service on the post or (ii) by direct, or (iii) by transfer or deputation of an officer already in the service of any State Government, or the Government of India."

3. It is according to Rule 9 that the post of Deputy Labour Commissioner is to be filled up from amongst the Labour Officer-cum-Conciliation Officer, statistical Officer, and Welfare Officer (Women). [...] In any case, even if the Rules have not been notified, the same can be taken as executive instructions.

⋮

6. The matter was heard on 4.10.1993. In view of the submissions made, it was though appropriate that a direction be given to the State to file a detailed reply specifically indicating whether Draft Rules have been approved by the Public Service Commission and given effect thereto or it is the stand of the State that the post of Statistical Officers are not at all to be considered for the purpose of promotion to the posts of Deputy Labour Commissioners. [...] It is a settled law that Draft Rules are no Rules in the presence of notified Rules. It is also clarified that Class I and II are redesignated as Group A and B respectively."

7. We have heard learned counsel for the parties and perused the relevant material referred to during the course of their submissions. The petitioner has challenged the appointment of respondent No. 3 to the post of Deputy Labour Commissioner primarily on two counts : (i) that while promoting the respondent No. 3, claim of the petitioner has not been considered; (ii) that Punjab Labour Service (Class I and II) Rules, 1955, having already been repealed and the Draft Rules approved by the Haryana Public Service Commission, entitle the petitioner for consideration in terms of Rule 9; and (iii) the Draft Rules having been approved by the Public Service Commission even if no notified can be taken as executive instructions and if construed so, the petitioner's case has merit.

8. Admittedly, under 1955 Rules there was no provision with the Labour Department for promotion to the post of Statistical Officer. It is on the advice of the Haryana Public Service Commission that avenue was created subject to the persons possessing the requisite qualifications and experience. [...] Even otherwise, as per decision in Jitender Kumar Verma's case (supra), Draft Rules cannot be equated with administrative instructions. Thus, finding no merit in any of the contentions raised by the counsel for the petitioner, the civil writ petition is dismissed. Parties to hear their own costs.

Table 4: Sample abstract and Judgment from the CIVILSUM test set (ID = 716).

Reference abstract (Tokens: 52)
Constitution of India, Article 300-A - Retirement Pension - Denial - The petitioner was previously employed before entering into the University and was receiving pension from their previous employer, therefore, the petitioner is eligible for pension from the University. **[Paras 4 and 12]**
Longformer summary Tokens: 142; ROUGE-1: 34.73; ROUGE-2: 15.76; ROUGE-L: 28.74
constitution of india, articles 14 and 300-a - university pension regulations, 1990, regulations 2 and 3(2) - pension - re-employment - discrimination - petitioner, who was appointed as a security guard in the university on 8.1.80 and had his services regularised and regular pay scale granted, sought pensionary benefits. however, the respondent university argued that the petitioner was not covered under the definition of re-employed as he was already receiving pension from his previous employer. the court held that the respondent's action was violative of articles 14, as the judgment of the single judge affirmed the entitlement of the petitioner to pension by the university. consequently, the university was directed to grant all retiral benefits from the date of superannuation. **[paras 9 to 14]**
Llama 2-chat summary Tokens: 112; ROUGE-1: 20.65; ROUGE-2: 5.23; ROUGE-L: 16.77
The five connected civil writ petitions were disposed of by the court, with the judgment in favor of the petitioner. The petitioner, who was a security guard at the University of Rajasthan, had been denied pensionary benefits on the ground that he was a re-employed pensioner. The court held that the petitioner was entitled to all retiral benefits as per the provisions of the 1990 Regulations, and the arrears of the retiral benefits were to be paid to the petitioner within 60 days. The court also directed that in case of non-payment of arrears, the petitioner would be entitled to interest on the due amount at the rate of 12% per annum.
FactorSum - Tail content guidance Tokens: 154; ROUGE-1: 31.77; ROUGE-2: 13.10; ROUGE-L: 23.53
Constitution of India, Articles 14 and 16 - Pension Regulations for University of Rajasthan, 1990, Regulation 3(b) - Retiral benefits - Re-employment - The petitioner, who was a regular employee of the respondent University, retired on attaining the age of superannuation and was governed by the Pension Regulations which were applicable to University Employees as per Register No. 2 subject to the provisions of proviso (a) to (d) and that the petitioner was not falling under any of the categories of the proviso - Therefore, the petitioner is entitled to all retiral benefits as per the 1990 Regulations. Constitution of India, Articles 14, 16 and 226 - Pension - Re-employment - The petitioner, who retired as a Security Guard in the University, is entitled to all retiral benefits from the date of his superannuation.
FactorSum - Lead content guidance Tokens: 127; ROUGE-1: 34.90; ROUGE-2: 10.88; ROUGE-L: 22.82
Constitution of India, Articles 14, 16 and 226 - Pension - Re-employment - The petitioner, who retired as a Security Guard in the University, is entitled to all retiral benefits from the date of his superannuation. Constitution of India, Articles 14 and 16 - Rajasthan Civil Services (Pension) Rules, 1990, Rule 3 - Retiral benefits - Re-employment - The petitioner, who was appointed as a Security Guard on 8.1.80, retired on 30.4.94 and was governed by the Pension Regulations applicable to University Employees as per Register No. 2 subject to the provisions of proviso (a) to (d) and was not falling under any of the categories of the proviso.

Table 5: Sample abstract and generated summaries from the CIVILSUM test set (ID = 2). FactorSum summaries includes the same sentences in different order. References to paragraphs in the source document are shown in blue color.

Reference abstract (Tokens: 94)
Under the Uttar Pradesh Public Services (Reservation for Scheduled Castes, Scheduled Tribes and other Backward Classes) Act, 1994, Rule 3(4) for Recruitment Eligibility, the State Engineering Services (Special Recruitment) Examination is conducted for the recruitment of SC and ST candidates.', 'In this process, it is reasonable to fix a minimum aggregate of 40% marks as the standard for evaluating candidate suitability.', 'It was established that the petitioner did not obtain the required minimum marks and was subsequently not called for an interview, therefore, there is no justification for intervention. **[Paras 7, 8, and 9]**
Longformer summary Tokens: 98; ROUGE-1: 58.95; ROUGE-2: 31.91; ROUGE-L: 43.16
constitution of india, article 16 - u.p. public services (reservation for scheduled castes, scheduled tribes and other backward classes) act, 1994, section 3(4) - interview - suitability - petitioners who appeared in state engineering services (special recruitment) examination, 1995 conducted by respondent no. 2 and were not called for interview - it was held that the petitioners were unsuitable for interview as they could not secure the 40% minimum marks required for judging their suitability. however, the criterion was not deemed arbitrary and did not violate any laws. hence, the petition was dismissed. **[paras 7, 8 and 9]**
Llama 2-chat summary Tokens: 109; ROUGE-1: 39.81; ROUGE-2: 16.67; ROUGE-L: 26.21
The petitioners, who were candidates in a special recruitment examination for S.C. and S.T. candidates, filed a writ petition against the respondents, claiming that they were not called for an interview due to the respondents' illegal fixing of 40% minimum marks in aggregate. The Court found that the requirement of securing 40% minimum marks in aggregate was not arbitrary and did not suffer from any legal infirmity. Therefore, the Court dismissed the writ petition, finding that the respondents did not commit any error in not calling the petitioners for an interview. The petitioners did not obtain the required 40% minimum marks in aggregate, making them ineligible for the interview.
FactorSum - Tail content guidance Tokens: 167; ROUGE-1: 47.49; ROUGE-2: 17.51; ROUGE-L: 21.0
T. candidates - The requirement of 40% marks in aggregate for assessing the suitability of a candidate for being called for interview is not illegal . **[Para 5]** Constitution of India, Articles 14, 16 and 226 - Appointment - Selection - Criteria of selection - The minimum standard for assessing the suitability of candidates is a minimum of 40% marks in aggregate - The fixing of the 40% minimum marks by respondent No. 2 does not suffer from any legal infirmity warranting interference by the Court . **[Para 3]** Constitution of India, Articles 14, 16 and 226 - Selection - Interview - Selection for the post of S . C. and S .T. candidates - Petitioners appeared in State Engineering Services (Special Recruitment) Examination, 1995 conducted by respondent No. 2 for recruitment of SC and ST candidates. T. candidates requires a minimum of 40% marks in aggregate.
FactorSum - Lead content guidance Tokens: 155; ROUGE-1: 54.88; ROUGE-2: 27.23; ROUGE-L: 26.05
Constitution of India, Articles 14, 16 and 226 - Selection - Interview - Selection for the post of S . C. and S .T. candidates - Petitioners appeared in State Engineering Services (Special Recruitment) Examination, 1995 conducted by respondent No. 2 for recruitment of SC and ST candidates. T. candidates - The requirement of 40% marks in aggregate for assessing the suitability of a candidate for being called for interview is not illegal . **[Para 5]** P. Public Services (Reservation for Scheduled Castes, Scheduled Tribes and other Backward Classes) Act, 1994, Section 4 - Recruitment - Interview - Post of Lecturer - Petitioners, who were appointed as Lecturers, challenged the appointment of Respondent No. 2 as Lecturer after obtaining 40% marks in aggregate. Constitution of India, Articles 14 and 16 - U .

Table 6: Sample abstract and generated summaries from the CIVILSUM test set (ID = 8). References to paragraphs in the source document are shown in blue color.