Contextors at L-SUMM: Retriever-Driven Multi-Generator Summarization

Pavithra Neelamegam, S. Jaya Nirmala
Department of Computer Science and Engineering
National Institute of Technology, Tiruchirappalli, Tamil Nadu, India
{406424001, sjaya}@nitt.edu

Abstract

Indian court judgments are very difficult to automatically summarize because of their length, complex legal reasoning and scattered important information. This paper outlines the methodology used for the Legal Summarization (L-SUMM) shared task at the JUST-NLP 2025 Workshop, which aims to provide abstractive summaries of roughly 500 words from English-language Indian court rulings that are logical, concise and factually accurate. The paper proposes a Retriever-Driven Multi-Generator Summarization framework that combines a semantic retriever with fine-tuned encoder–decoder models BART, Pegasus and LED to enhance legal document summarization. This pipeline uses cosine similarity analysis to improve summary faithfulness, cross-model validation to guarantee factual consistency and iterative retrieval expansion to choose relevant text chunks in order to address document length and reduce hallucinations. Despite being limited to 400–500 words, the generated summaries successfully convey legal reasoning. Our team *Contextors* achieved an average score of 22.51, ranking 4th out of 9 in the L-SUMM shared task leaderboard, demonstrating the efficacy of Retriever-Driven Multi-Generator Summarization approach, which improves transparency, accessibility, and effective understanding of legal documents. This method shows excellent content coverage and coherence when assessed using ROUGE-2, ROUGE-L, and BLEU criteria.

1 Introduction

Finding an important information in lengthy, complex and unstructured court case judgments can be difficult for legal professionals. These documents are often hundreds of words long and need a lot of time and effort to read and understand. Such documents need to be manually summarized, and it is a costlier and time-consuming process. It also requires expert legal knowledge. To address the challenge of processing lengthy legal judgments, the JUST-NLP 2025 Workshop conducted the Shared Task on Legal Summarization (L-SUMM) .

This work promotes the creation of AI-powered solutions that can automatically produce abstractive summaries of Indian court rulings. Abstractive summarizing creates new, coherent text that translates, condenses and rephrases complicated legal jargon into instructive summaries of about 500 words that capture the core of the ruling, in contrast to extractive summary, which chooses preexisting sentences from a document.

This approach uses transformer-based encoder–decoder designs like BART, T5, Pegasus and LED, which are further enhanced by retriever models based on cosine similarity analysis for semantic chunk selection. These models capture the unique terminology, reasoning processes and discourse structures of Indian court decisions by fine-tuning on domain-specific legal data.

Submissions to the shared work are evaluated using standard relevance metrics, such as ROUGE-2, ROUGE-L and BLEU, which evaluate the quality, fluency and overlap of the generated summaries with human references.

2 Related Work

For Indian legal papers, (Ghosh et al., 2022) suggested a text normalization method that standarizes reference styles, legal jargon and acronyms before fine tuning generic models. But, it fails to capture long-document connections and hierarchical structures. (Deroy et al., 2024) examined the LLMs for summarizing judgments by contrasting extractive and abstractive approaches but suffered from factual errors and limited citations awareness. (Santosh et al., 2024) presented LexAbSumm, an aspect-based framework improving interpretability, but has a trouble handling inter aspect interdependence.

Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025), pages 107–112

BART (Lewis et al., 2019) achieves strong summarization quality but is constrained by its limited 1K token input window. Pegasus (Zhang et al., 2020) better aligns with summarization tasks. However, its context window and domain generality limit applicability to legal reasoning tasks. Furthermore, after tuning Pegasus on domain-specific corpora, Legal Pegasus(Sharma and Singh, 2024) enhanced factual consistency but failed to sustain complex citation relations.To address long-document contexts, BigBird (Zaheer et al., 2020) proposed a sparse-attention mechanism for efficient scaling and LED (Beltagy et al., 2020) extended this idea to handle up to 16K tokens, enhancing length coverage, but both the approaches struggle in capturing hierarchical legal semantics.

Several transformer-based summarizers (BART, T5, Pegasus) were merged in ensemble frameworks (Albayati et al., 2025) to improve factuality. However, it provides only a minimal improvement at the expense of significant computational complexity. More recently, RAG models (Ajay Mukund and Easwarakumar, 2025) have integrated external knowledge retrieval with generative summarization, however, still face issues related to retrieval precision, latency, and maintaining structural coherence in lengthy legal texts. Further research is required to capture the facts, issues, laws and other crucial components.

Our approach differs from previous legal summarization methods by introducing a retrieval-guided, multi-generator framework. It first uses InLegalBERT to select the most relevant judgment segments, then combines outputs from multiple fine-tuned abstractive models instead of relying on a single generator. A faithfulness-based semantic alignment score is finally used to choose the most accurate summary, resulting in a retrieval-aware and fact-faithful summarization method not present in prior work.

3 Dataset Description

The InLSum (Indian Legal Summarization) dataset, used for this shared task, includes 1200 training samples, 200 validation samples and 400 test samples of Indian court rulings that are accompanied with abstractive summaries produced by legal experts. The InLSum dataset is provided in JSONL format.The training dataset contains judgment and reference summary files. The validation and test dataset contains only judgments.

4 Model Description

4.1 Fine-Tuned Pre-trained Models

A variety of pre-trained sequence-to-sequence models, such as BART, Legal-Pegasus, T5 and LED, has been finetuned using InLSumm dataset. Every judgment and summary is preprocessed to remove extraneous text, page numbers and repeating blocks, so that the models can focus on the most important information. During the training phase, 90% of dataset is used for training and 10% of dataset is used for validation purpose. The models are adjusted to produce an abstract summary of 400–500 words. ROUGE-2, ROUGE-L and BLEU are used to evaluate overall quality of the summarization.

4.2 Legal Ensemble Summarization Framework

This work presents an ensemble based abstractive summarization framework for legal judgments that integrates multiple fine tuned transformer architectures. Each judgment is preprocessed to remove unnecessary components such as case identifiers, citations, and formatting patterns. In the ensemble configuration, each model generates a summary for the same input independently. BART and Pegasus, BART and LED, and Pegasus and LED are pairwise ensembled to take advantage of the strengths of BART's fluency, LED's sparse attention for lengthy sequences, and Pegasus's summary pre-training. Clarity, sentence versatility and legal relevance are further improved by a 3-way hybrid ensemble that combines BART, Pegasus, and LED.

Semantic ranking is used for selection using InLegalBERT(Sharma and Singh, 2024). Sentence embeddings will be calculated for both candidate summaries and the initial judgment. The final result is then determined by selecting the summary with the highest cosine similarity. In order to offer a thorough assessment of summarization quality, the final evaluation employs ROUGE-2, ROUGE-L, and BLEU, which examine text fluency, structure similarity, and content accuracy.

4.3 Retriever-Driven Multi-Generator Summarization

The proposed Retriever-Driven Multi-Generator Summarization framework, offers a novel method for producing concise, logical, and factually consistent summaries of complicated legal rulings. Semantic retrieval, multi-generator fine-tuning, and

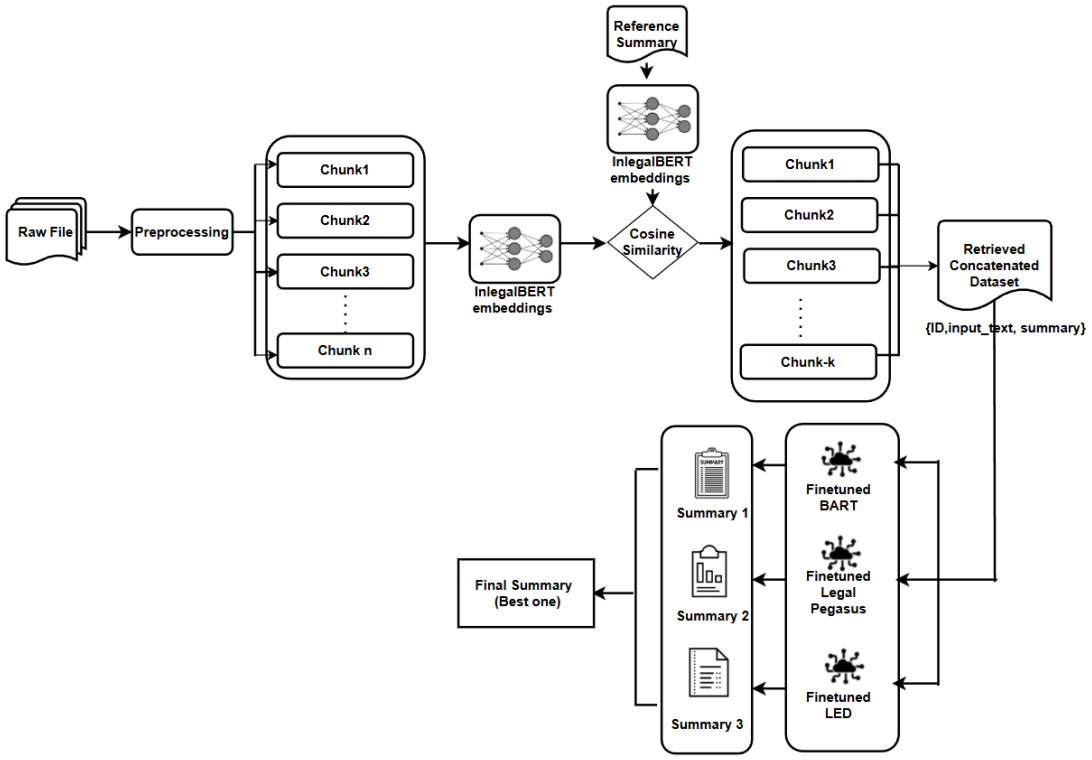


Figure 1: Retriever-Driven Multi-Generator Summarization framework

automatic ensemble selection are the three main components of this framework.As shown in Figure 1, this proposed framework combines retrieval and generation modules.

After preprocessing, to maintain compatibility with transformer input restrictions, each file is split into reasonable chunks (chunk_1, chunk_2,..., chunk_n) due to the length of legal documents. Judgments are split into 8-sentence chunks, embedded using 768-D vectors, and ranked by cosine similarity against the full-judgment embedding to measure semantic relevance. The system retrieves the top 8 most relevant chunks for summary generation. The obtained dataset is composed of tuples of ID, input_text, summary. By ensuring that only parts of the document that are most contextually aligned contribute to model fine-tuning, this retrieval procedure improves efficiency and factual grounding.

90% of training data and 10% of validation data are created from the retrieved and concatenated dataset. This dataset is used to independently train three transformer-based summarization models LED, Legal Pegasus and BART.

Each finetuned model independently produces a summary for the same input document in the ensemble setting. The embeddings obtained from the validation judgment datatsets are then compared to these outputs (Summary1, Summary2, and Summary3) using cosine similarity. In order to ensure contextual accuracy and relevance, the summary with the best semantic similarity score is chosen as the final summary. The chosen final summaries are statistically assessed for n-gram overlap, fluency, and informativeness using the average of the ROUGE-2, ROUGE-L, BLEU scores.

Training Phase: To find the most important textual parts, the system uses chunk–summary similarity in a supervised retrieval-enhanced training configuration. From these retrieved pieces, the summarizing model is subsequently refined to provide summaries.

Inference Phase: Since gold summaries are not available, the retrieval process runs unsupervised during inference. In order to extract the most semantically representative chunks, the model calculates the mean embedding of the input judgment. These chunks are then supplied to the refined summarizer to provide abstractive summaries.

All preprocessing scripts, codes used in this work are publicly available online. [1]

[1] https://github.com/pavithraneelamegam/Summarization

Table 1: Performance comparison of fine-tuned and ensemble models using ROUGE and BLEU metrics.

Model	ROUGE-2	ROUGE-L	BLEU	AVG
Fine-Tuned Models				
Fine-Tuned BART	17.79	25.32	14.53	19.21
Fine-Tuned Legal Pegasus	17.64	25.12	13.66	18.81
Fine-Tuned LED	16.74	24.02	12.51	17.76
Fine-Tuned T5	16.08	23.02	11.03	16.71
Ensemble Models				
Fine-Tuned (BART and Pegasus)	23.02	25.20	15.47	21.23
Fine-Tuned (BART and LED)	23.40	25.40	16.06	21.62
Fine-Tuned (BART, Pegasus and LED)	23.62	25.60	16.69	21.97
Retriever-Driven Multi-Generator Summarization	**25.13**	**25.59**	**16.8**	**22.51**

5 Evaluation Metrics

5.1 ROUGE Scores

Automatic summarization evaluation is the main application of the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric. By computing n-gram recall, it determines how much the generated result summaries and reference summaries coincide. Equation 1 represents ROUGE-1 (unigram).

$$\mathrm{ROUGE}_1 = \frac{\sum_{\text{unigram}\in\text{reference}} \text{Count}_{\text{match}}(\text{unigram})}{\sum_{\text{unigram}\in\text{reference}} \text{Count}(\text{unigram})} \tag{1}$$

5.2 BLEU Score

BLEU (Bilingual Evaluation Understudy) captures sufficiency, fidelity, and fluency by measuring the amount of overlap of n-grams among a reference phrase and a candidate (hypothesis).

6 Results and Discussion

Table 1 presents the comparative performance of individual and ensemble summarization models on the InLSum dataset. BART outperformed the other finetuned models (AVG = 19.21), closely followed by Legal Pegasus (AVG = 18.81). This indicates that both architectures effectively model hierarchical dependencies and domain-specific semantics in legal documents. LED showed slightly lower performance (AVG = 17.76), likely due to the truncation of lengthy case texts and difficulty in handling factual segments.

The BART and LED combination showed a notable improvement (AVG = 21.62) because LED's contextual comprehension supports BART's improved surface realization and syntax fluency. The triple-model ensembles (BART, Pegasus and LED) produced the highest average (AVG = 21.97) , suggesting that integrating different attention and producing mechanisms results in more balanced summaries with enhanced coverage and factual coherence.

Outperforming all other models, the proposed Retriever-Driven Multi-Generator Summarization framework produced the best overall results (ROUGE-L = 25.59, BLEU = 16.8, AVG = 22.51). Each generator can now concentrate on the most semantically relevant context pieces prior to generation due to the retrieval-enhanced architecture. It generates summaries that are context-aware and factually consistent, better capturing interdependent facts and fine-grained legal reasoning than single models.

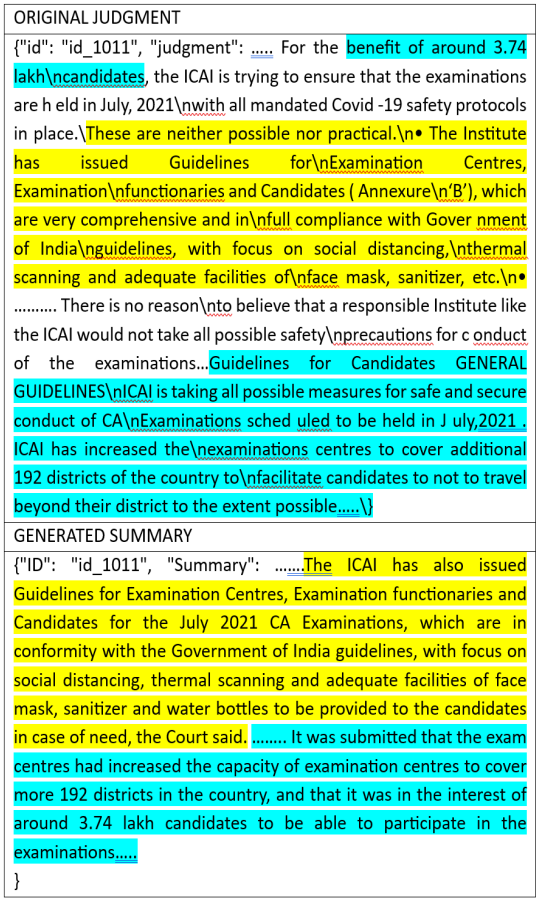
6.1 Qualitative Analysis

The generated summaries generally capture the main storyline of judicial decisions, including the key dispute, parties, and major facts. This shows that the model handles high-level narrative extraction well. However, it often misses important legal details such as statutes, procedural steps, and the Court's reasoning and may overlook subtle facts or introduce errors. As a result, the summaries are coherent but not fully accurate from a legal perspective.

These challenges show why abstractive summarization in law is difficult. Judgments are complex, detailed, and heavily structured around legal reasoning. Automated systems often struggle to pre-

serve the logical flow, handle layered arguments, or capture references to past cases. Therefore, even though the summaries offer a good general narrative, they still require human review and legal expertise to ensure accuracy and reliability.

In the first example, as shown in Figure 2 the model captures the main theme well, correctly summarizing the dispute over conducting exams with COVID-19 safety measures. However, it misses key legal details such as the statutory basis, opt-out rules, and specific safety guidelines making the summary less accurate and complete from a legal perspective.


ORIGINAL JUDGMENT

{"id": "id_1011", "judgment": ..... For the benefit of around 3.74 lakh\ncandidates, the ICAI is trying to ensure that the examinations are h eld in July, 2021\nwith all mandated Covid -19 safety protocols in place.\These are neither possible nor practical.\n• The Institute has issued Guidelines for\nExamination Centres, Examination\nfunctionaries and Candidates (Annexure\n'B'), which are very comprehensive and in\nfull compliance with Gover nment of India\nguidelines, with focus on social distancing,\nthermal scanning and adequate facilities of\nface mask, sanitizer, etc.\n• ..... There is no reason\nto believe that a responsible Institute like the ICAI would not take all possible safety\nprecautions for c onduct of the examinations...Guidelines for Candidates GENERAL GUIDELINES\nICAI is taking all possible measures for safe and secure conduct of CA\nExaminations sched uled to be held in J uly,2021 . ICAI has increased the\nexaminations centres to cover additional 192 districts of the country to\nfacilitate candidates to not to travel beyond their district to the extent possible.....\}

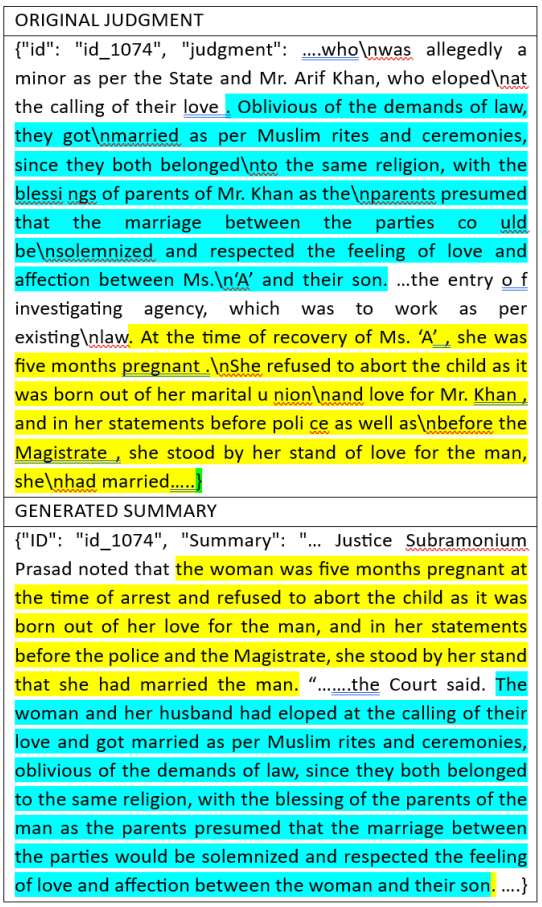
GENERATED SUMMARY

{"ID": "id_1011", "Summary": .....The ICAI has also issued Guidelines for Examination Centres, Examination functionaries and Candidates for the July 2021 CA Examinations, which are in conformity with the Government of India guidelines, with focus on social distancing, thermal scanning and adequate facilities of face mask, sanitizer and water bottles to be provided to the candidates in case of need, the Court said. ..... It was submitted that the exam centres had increased the capacity of examination centres to cover more 192 districts in the country, and that it was in the interest of around 3.74 lakh candidates to be able to participate in the examinations.....

}

Figure 2: original judgment text vs generated summary(sample 1)

In the second example as shown in Figure 3, the model captures the main narrative and human aspects of the judgment well, but it misses key legal reasoning, overlooks the age-related dispute, and includes minor inaccuracies. As a result, the summary conveys the story but lacks the legal precision needed to fully reflect the Court's reasoning.


ORIGINAL JUDGMENT

{"id": "id_1074", "judgment": ....who\nwas allegedly a minor as per the State and Mr. Arif Khan, who eloped\nat the calling of their love . Oblivious of the demands of law, they got\nmarried as per Muslim rites and ceremonies, since they both belonged\nto the same religion, with the blessi ngs of parents of Mr. Khan as the\nparents presumed that the marriage between the parties co uld be\nsolemnized and respected the feeling of love and affection between Ms.\n'A' and their son. ...the entry o f investigating agency, which was to work as per existing\nlaw. At the time of recovery of Ms. 'A' , she was five months pregnant .\nShe refused to abort the child as it was born out of her marital u nion\nand love for Mr. Khan , and in her statements before poli ce as well as\nbefore the Magistrate , she stood by her stand of love for the man, she\nhad married.....}

GENERATED SUMMARY

{"ID": "id_1074", "Summary": "... Justice Subramonium Prasad noted that the woman was five months pregnant at the time of arrest and refused to abort the child as it was born out of her love for the man, and in her statements before the police and the Magistrate, she stood by her stand that she had married the man. "....the Court said. The woman and her husband had eloped at the calling of their love and got married as per Muslim rites and ceremonies, oblivious of the demands of law, since they both belonged to the same religion, with the blessing of the parents of the man as the parents presumed that the marriage between the parties would be solemnized and respected the feeling of love and affection between the woman and their son. ....}

Figure 3: original judgment text vs generated summary(sample 2)

7 Conclusion

The JustNLP Shared Task underscores the potential of NLP techniques in addressing the challenges of abstractive summarization for complex legal documents. In this study, the proposed Retriever-Driven Multi-Generator Summarization framework that integrates multiple fine-tuned transformer models with a retrieval based preprocessing pipeline to generate coherent, legally faithful and semantically rich summaries. This ensemble architecture outperforms individual fine-tuned models, yielding an overall improvement of +2-3 ROUGE points over evaluation criteria. To further enhance summary reliability and interpretability, future research will concentrate on hierarchical retrieval, fidelity optimization based on reinforcement learning, and legal provision-aware structural modeling.

Limitations

There is a trade-off between faithfulness and abstraction in the suggested Retriever-Driven Multi-Generator Summarization framework. Accuracy

at the sentence level cannot be guaranteed by existing embedding-based checks. The quality of retrieval has a significant impact on the proposed framework's performance as well. The model treats judgments as plain text rather than acknowledging their hierarchical structure, rhetorical functions and argument flow, all of which are crucial for creating logical and legally accurate summaries. Lastly, although retrieval introduces only moderate overhead, the use of multiple large transformer models makes the overall pipeline computationally intensive. A precise complexity analysis is difficult due to architecture-dependent embedding and generation costs. Therefore, we plan to include detailed profiling and computational analysis in future work.

Acknowledgements

This work was funded by the Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India, under Grant No. CRG/2023/007683.

References

S Ajay Mukund and KS Easwarakumar. 2025. Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation. *Symmetry*, 17(5):633.

Maha Ahmed Abdullah Albayati, Kürşat Mustafa Karaoğlan, and Oğuz Findik. 2025. Towards efficient multi-legal document summarization: An ensemble approach for turkish law. *Engineering Science and Technology, an International Journal*, 70:102138.

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

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2024. Applicability of large language models and generative models for legal case judgement summarization.

Satyajit Ghosh, Mousumi Dutta, and Tanaya Das. 2022. Indian legal text summarization: A text normalization-based approach. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–4.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

T. Y. S. S Santosh, Mahmoud Aly, and Matthias Grabmair. 2024. Lexabsumm: Aspect-based summarization of legal decisions. *Preprint*, arXiv:2404.00594.

Saloni Sharma and Piyush Pratap Singh. 2024. Domain-specific summarization: Optimizing inlegalbert for indian judgment reports.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.