The Role of Context in Sequential Sentence Classification for Long Documents

Anonymous ACL submission

Abstract

Sequential sentence classification extends traditional classification by incorporating broader context. However, state-of-the-art approaches face two major challenges in long documents: pretrained language models struggle with input-length constraints, while proposed hierarchical models often introduce irrelevant content. To address these limitations, we propose a document-level retrieval approach that extracts only the most relevant context. Specifically, we introduce two heuristic strategies: **Sequential**, which captures local information, and **Selective**, which retrieves the most semantically similar sentences. Experiments on legal domain datasets show that both heuristics improve performance. Sequential heuristics outperform hierarchical models on two out of three datasets, demonstrating the benefits of targeted context.

1 Introduction

Sequential sentence classification (SSC) is the task of categorizing sentences based on their semantic role within a document. Since a sentence's meaning is often shaped by its surrounding context, SSC is particularly useful in structured texts such as legal cases. Identifying key rhetorical components (e.g., preamble, issue, or analysis; see Figure 1) benefits downstream tasks such as information retrieval (Neves et al., 2019; Safder and Hassan, 2019) and document summarization (Kalamkar et al., 2022; Muhammed et al., 2024).

State-of-the-art hierarchical models have achieved strong performance on SSC by processing entire document sequences at once, thereby capturing a broader context (Jin and Szolovits, 2018; Brack et al., 2021; Kalamkar et al., 2022). However, we make the assumption that focusing on all sentences may not always be necessary, as this can introduce noise from irrelevant content (Shi et al., 2023). Additionally, pretrained language models (PLMs) remain constrained by input-length limitations (Warner et al., 2024), even with advancements in large language models (LLMs) (BehnamGhader et al., 2024).

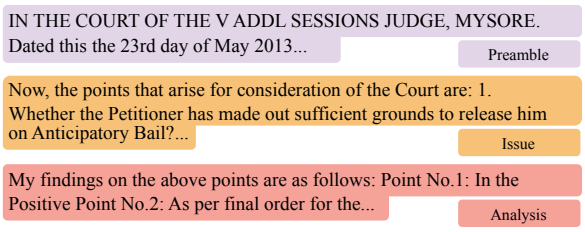


Figure 1: A segment of a legal document with sentences labeled by their function.

Recent studies have explored strategies for retrieving relevant information at the document level (Amalvy et al., 2023b; Lan et al., 2024). Yet, to our knowledge, no existing work has explicitly investigated how to retrieve the most relevant context to optimize PLM performance in SSC.

In this paper, our contributions are twofold: (1) analyzing the role of context in SSC by introducing two heuristic retrieval strategies—*Sequential*, which leverages local information around each sentence, and *Selective*, which retrieves the most semantically similar sentences at the document level—and (2) demonstrating how these strategies enhance PLMs by providing more relevant context.

We evaluate on document-level datasets in the legal domain, the primary benchmark for SSC tasks. To foster transparency and reproducibility, we release our code under an open-source license[1].

2 Related Work

2.1 Input Sequence Constraints in PLMs

Encoder-only models such as BERT (Devlin et al., 2019) offer a strong tradeoff between size and performance, making them a compelling alternative to larger decoder-based architectures for classification tasks. However, the quadratic complexity of self-attention in vanilla Transformer models lim-

[1] https://anonymous.4open.science/r/ACL-2025-4BE2

its their effective input length, posing challenges for processing long documents. To mitigate this, sparse attention mechanisms have been introduced to reduce computational costs (Zaheer et al., 2020; Wang et al., 2020; Beltagy et al., 2020; Choromanski et al., 2020). While these methods extend the context range, they still fall short of fully resolving the limitations of long-text processing (Warner et al., 2024; Nussbaum et al., 2025).

2.2 SSC for Long Documents

Early work on SSC focused on hierarchical models to incorporate broader context into sentence representations. Hierarchical Sequential Labeling Network (HSLN) was among the first frameworks to process full-document sequences for contextualized representations (Jin and Szolovits, 2018; Shang et al., 2021; Brack et al., 2021; Kalamkar et al., 2022). More recent studies have explored refined learning strategies: T.y.s.s. et al. (2024) applied contrastive and prototypical learning to enhance sentence representations by leveraging semantic similarities, while Santosh et al. (2024) introduced a hierarchical curriculum learning framework to progressively improve the model's ability to distinguish rhetorical labels at different levels of granularity.

While these studies have primarily focused on improving HSLN, our work addresses a different challenge: overcoming input-length constraints in PLMs by retrieving only the most relevant context, thereby reducing noise and improving efficiency in SSC.

3 Context Retrieval

To investigate the role of context in enhancing PLM performance, we define two types of heuristics: **Sequential** and **Selective**. These heuristics determine which sentences should be incorporated into the model's input and are inspired by prior research on contextual enrichment in the era of LLMs (Amalvy et al., 2023a; Wang et al., 2024; Nussbaum et al., 2025).

Sequential Heuristics extract context from sentences adjacent to the target sentence within the same document. We consider three strategies:

- **Before**: Selects the k sentences immediately preceding the target sentence.
- **After**: Selects the k sentences immediately following the target sentence.
- **Surrounding**: Selects $\frac{k}{2}$ sentences before and after the target sentence.

Selective Heuristics unlike sequential strategies, selective heuristics retrieve sentences from anywhere in the document, independent of their position relative to the target sentence. We explore three selection techniques:

- **Random**: Randomly selects k sentences from the entire document.
- **BM25**: Retrieves the k most relevant sentences using BM25 (Trotman et al., 2014), a ranking function that scores sentences based on a term frequency-inverse document frequency (TF-IDF) weighting scheme. BM25 is widely used in information retrieval for lexical relevance ranking.
- **Sentence-BERT**: Selects the k semantically closest sentences to the target sentence using Sentence-BERT embeddings (Reimers and Gurevych, 2019), which capture sentence-level similarity via a fine-tuned siamese BERT network.

Given computational constraints, we limit our analysis to $k = 6$. Notably, selective heuristics may retrieve sentences that are also included in the sequential context since they operate over the full document. Table 3 in the Appendix provides illustrative examples.

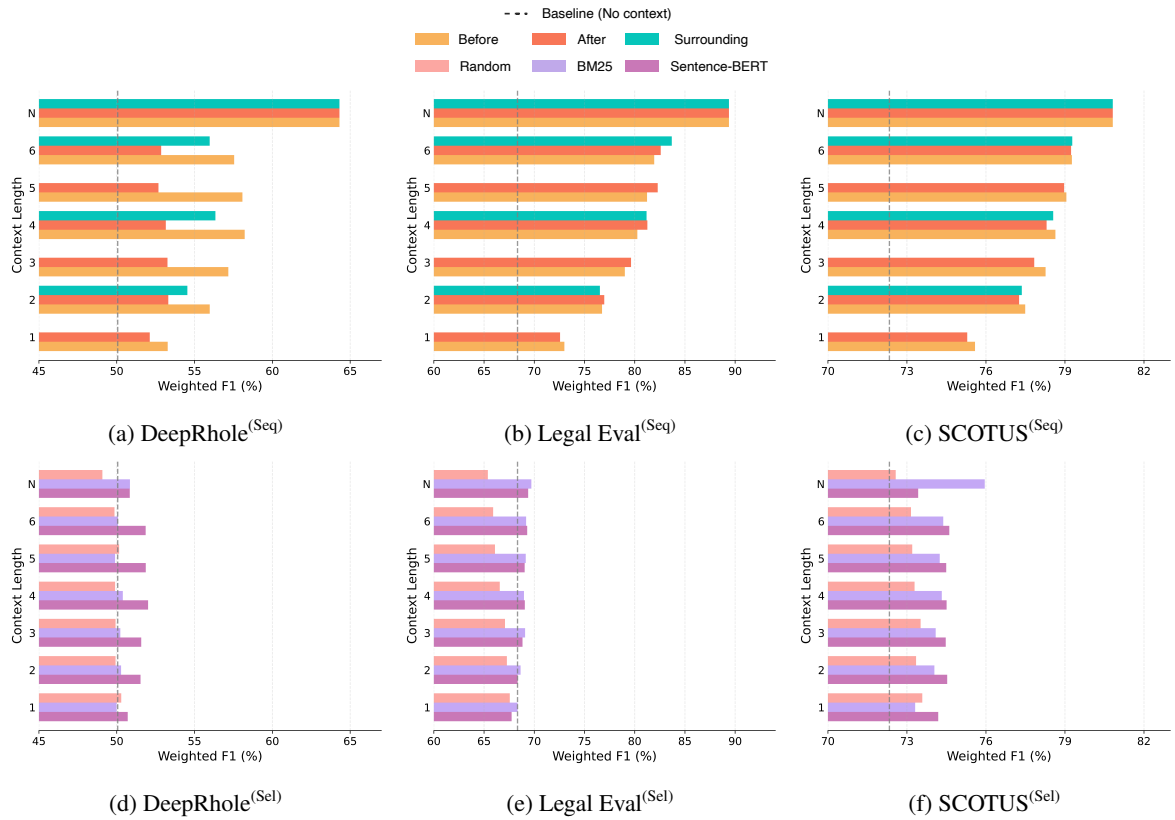
Sentence Ordering We further investigate whether the order of retrieved sentences impacts SSC performance. Inspired by NAREOR (Gangal et al., 2022), which explores sentence reordering to analyze narrative coherence in storytelling, we examine whether maintaining full document sentences ($k = N$) while altering their order affects performance.

To evaluate this, we use our heuristics. In Sequential, we retain the original human-written order to preserve logical flow. In Selective, we reorder sentences based on their relevance to the target sentence while ensuring that all remain included for a fair comparison.

4 Experimental protocol

4.1 Datasets

Our experiments focus on the legal domain, as it is the only domain with datasets annotated at the document level. We utilize three datasets:



(a) DeepRhole$^{(Seq)}$ (b) Legal Eval$^{(Seq)}$ (c) SCOTUS$^{(Seq)}$

(d) DeepRhole$^{(Sel)}$ (e) Legal Eval$^{(Sel)}$ (f) SCOTUS$^{(Sel)}$

Figure 2: Weighted F1 scores for different context lengths k across three datasets. The top row (a, b, c) presents results using Sequential context$^{(Seq)}$, while the bottom row (d, e, f) represents Selective context$^{(Sel)}$. $k = N$ indicates that the full document is used to address the sentence ordering question. We set k as an even number for Surrounding heuristic to ensure comparability in context length with other ones.

(i) DeepRhole (Bhattacharya et al., 2023), (ii) Legal-Eval (Kalamkar et al., 2022), and (iii) SCOTUS (Lavissière and Bonnard, 2024), derived from Indian and U.S. legal judgments. DeepRhole contains 7 rhetorical role labels, while the others have 13 each. A detailed dataset description is provided in Appendix A[2].

In contrast, other existing datasets (Dernoncourt et al., 2017; Gonçalves et al., 2020; Lan et al., 2024) primarily focus on scientific and biomedical abstracts, averaging only 10 sentences per sample. Their lack of document-level annotations makes them unsuitable for this study. For evaluation, we report the weighted F1-score.

4.2 SSC Model for Context Analysis

Our analysis builds upon the hierarchical HSLN model (Brack et al., 2021), with two minor modifications: (1) Motivated by ablation studies (Jin and Szolovits, 2018; Chen et al., 2023), which identified the contextual sentence enrichment layer as HSLN's primary driver of effectiveness, we removed the conditional random field (CRF) layer, and (2) We optimize only over the target sentence, enriched with context selected by our heuristics. Further architectural details, including our refinements, are provided in Appendix B. All results are averaged over three runs for robustness.

[2]All datasets were split at the document level into 80% training, 10% validation, and 10% test sets.

5 Results

5.1 Context Analysis

Figure 2 demonstrates that incorporating contextual sentences consistently improves classification performance across all datasets, regardless of the heuristic applied. This confirms the importance of effective context selection in SSC.

Sequential Heuristics systematically improve classification as more sentences are included. In Legal-Eval and SCOTUS, the Surrounding heuristic achieves the highest F1 score (83.6% and 79.2% at $k = 6$, respectively). However, in DeepRhole, the Before heuristic performs best, reaching 58.2%. A closer examination reveals that 71% of correctly assigned labels are shared across sequential heuris-

Model	Seq	DeepRhole	Legal Eval	SCOTUS
BERT (baseline)	512	52.23	69.74	75.58
+ Before		**67.18**†	78.41†	79.74†
+ After		56.72†	**79.74**†	**81.34**†
+ Surrounding		62.87†	77.27†	75.47
+ Random		46.86	67.05	74.70
+ BM25		51.59	69.43	75.96
+ Sentence-BERT		52.23	68.98	76.24
Nomic-BERT (baseline)	2048	50.32	68.90	75.50
+ Before		**67.89**†	80.54†	81.12†
+ After		57.75†	**81.11**†	**81.32**†
+ Surrounding		65.51†	78.20†	80.81†
+ Random		51.61	68.43	75.73
+ BM25		53.90	70.82‡	77.06†
+ Sentence-BERT		54.02‡	70.76‡	77.17‡
BERT-HSLN (SOTA)	$512 \times N$	54.45	93.06	79.66

Table 1: Performance of PLMs using the best configuration observed in context analysis for $k \leq 6$ for each heuristic. Bold values represent the best improvement over the baseline (w/o context), while underlined values indicate the second-best. BERT-HSLN is the SOTA for the SSC task. Markers † and ‡ denote statistical significance over the baseline at $p = 0.05$ and $p = 0.01$, respectively.

tics, suggesting that performance converges regardless of the specific choice.

In contrast, **Selective Heuristics** yield marginal gains, with BM25 being the most effective, reaching $\approx 74\%$ F1 in SCOTUS when $k \leq 6$.
The limited effectiveness of those heuristics could be attributed to two factors: (1) When documents lack semantically similar sentences, heuristics retrieve unrelated ones, adding noise (as observed in DeepRhole), and (2) Heuristics are most effective when retrieved sentences share the same target label (Figure 3 in the Appendix).

At $k = N$, the **Sentence Ordering** experiment confirms that SSC is sensitive to how context is structured—with the highest scores observed when the document's logical flow is preserved. Conversely, reordering sentences using Selective heuristics suggests that taking the full document may not be necessary; instead, prioritizing only the most relevant ones yields competitive performance.

5.2 Context Enrichment for PLMs

To examine how PLMs benefit from contextual enrichment[3], we conduct experiments with BERT (Devlin et al., 2019) and the recently introduced Nomic-BERT (Nussbaum et al., 2025), as shown in Table 1.

Our results indicate that Sequential heuristics typically yield the largest improvements, significantly outperforming the no-context baseline. Notably, they outperform the state-of-the-art BERT-HSLN[4], which processes the entire document at once for DeepRhole and SCOTUS.
We attribute the substantial improvement, particularly in DeepRhole, to two key factors: (1) The dataset has fewer rhetorical labels compared to others, and (2) From a statistical point of view, on average, a new rhetorical label persists for approximately 8.56 sentences before transitioning to another label. As a result, fully hierarchical models like BERT-HSLN, which process broader document segments, may struggle with these shifts, leading to a loss of important contextual information[5][6].

However, Legal-Eval remains challenging, as these PLMs have not yet matched SOTA performance. A plausible explanation is its higher label complexity, making it difficult for small models like BERT to achieve strong discrimination, as noted in SCOTUS annotation guidelines (Lavissière and Bonnard, 2024).
Additional results with RoBERTa (Liu et al., 2019), LegalBERT (Chalkidis et al., 2020), and Longformer (Beltagy et al., 2020) are provided in Appendix C.

6 Conclusion and Future Work

In this study, we investigated how the role of context affects the SSC task in long legal documents. Our findings reveal that sequential context heuristics, which preserve the flow of text, systematically lead to stronger performance gains than selective context. Moreover, enriching PLMs such as BERT with useful context yielded significant improvements over hierarchical models that process entire documents. Future work should give priority to (1) expanding the study to the corpus level, where multi-document contexts will be explored, and (2) refining selective heuristics to extract high-quality context without increasing noise.

[3]Context sentences were integrated with the target sentence into the PLM input while maintaining the natural human order for sequential heuristics.

[4]For a fair comparison, we compare against the original model, which does not include our modifications introduced in context analysis.

[5]Segment refers to consecutive annotation units (sentences) that share the same label within a document.

[6]The statistics were computed based on our analysis of the corpus.

7 Limitations

While this study demonstrates the benefits of contextual information for SSC, few limitations must be considered:

- We purposefully kept the heuristics basic, as our focus is not on peak performance. Nonetheless, more sophisticated approaches may yield higher scores than what we present.
- We have focused in our experiments on a single document. In practice, integrating the context of several documents could potentially offer richer information for selective heuristics.
- We cannot reject the hypothesis that our findings about the utility of context may not be universally generalizable across other tasks. Our analysis centered on legal datasets, and thus further research is needed to determine whether similar gains would arise in other settings.

8 Ethical Statement

This work fully complies with the ACL Ethics Policy. We declare that there are no ethical issues in this paper, to the best of our knowledge.

References

Arthur Amalvy, Vincent Labatut, and Richard Dufour. 2023a. Learning to rank context for named entity recognition using a synthetic dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10372–10382, Singapore. Association for Computational Linguistics.

Arthur Amalvy, Vincent Labatut, and Richard Dufour. 2023b. The role of global and local context in named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 714–722, Toronto, Canada. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

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

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2023. Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. *Artificial Intelligence and Law*, pages 1–38.

A Brack, A Hoppe, P Buschermöhle, and R Ewerth. 2021. Sequential sentence classification in research papers using cross-domain multi-task learning. corr. *arXiv preprint arXiv:2102.06008*.

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.

Yu Chen, You Zhang, Jin Wang, and Xuejie Zhang. 2023. YNU-HPCC at SemEval-2023 task 6: LEGAL-BERT based hierarchical BiLSTM with CRF for rhetorical roles prediction. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2075–2081, Toronto, Canada. Association for Computational Linguistics.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell, and Adrian Weller. 2020. Rethinking attention with performers. *CoRR*, abs/2009.14794.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neural networks for joint sentence classification in medical paper abstracts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 694–700, Valencia, Spain. Association for Computational Linguistics.

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

Varun Gangal, Steven Y. Feng, Malihe Alikhani, Teruko Mitamura, and Eduard Hovy. 2022. Nareor: The narrative reordering problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10645–10653.

Sérgio Gonçalves, Paulo Cortez, and Sérgio Moro. 2020. A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Computing and Applications*, 32(11):6793–6807.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109, Brussels, Belgium. Association for Computational Linguistics.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022. Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Mengfei Lan, Lecheng Zheng, Shufan Ming, and Halil Kilicoglu. 2024. Multi-label sequential sentence classification via large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16086–16104, Miami, Florida, USA. Association for Computational Linguistics.

Mary C. Lavissière and Warren Bonnard. 2024. Who's really got the right moves? analyzing recommendations for writing american judicial opinions. *Languages*, 9(4).

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

Akheel Muhammed, Hamna Muslihuddeen, Shalaka Sankar, and M Anand Kumar. 2024. Impact of rhetorical roles in abstractive legal document summarization. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE.

Mariana Neves, Daniel Butzke, and Barbara Grune. 2019. Evaluation of scientific elements for text similarity in biomedical publications. In *Proceedings of the 6th Workshop on Argument Mining*, pages 124–135, Florence, Italy. Association for Computational Linguistics.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2025. Nomic embed: Training a reproducible long context text embedder.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Iqra Safder and Saeed-Ul Hassan. 2019. Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. *Scientometrics*, 119:257–277.

T.y.s.s Santosh, Apolline Isaia, Shiyu Hong, and Matthias Grabmair. 2024. HiCuLR: Hierarchical curriculum learning for rhetorical role labeling of legal documents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7357–7364, Miami, Florida, USA. Association for Computational Linguistics.

Xichen Shang, Qianli Ma, Zhenxi Lin, Jiangyue Yan, and Zipeng Chen. 2021. A span-based dynamic local attention model for sequential sentence classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 198–203, Online. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

Santosh T.y.s.s., Hassan Sarwat, Ahmed Mohamed Abdelaal Abdou, and Matthias Grabmair. 2024. Mind your neighbours: Leveraging analogous instances for rhetorical role labeling for legal documents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11296–11306, Torino, Italia. ELRA and ICCL.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang,

Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

A Dataset

We experiment on three SSC datasets:

(i) DeepRhole (Bhattacharya et al., 2023) consists of 50 judgments from the Supreme Court of India, spanning five legal domains. It includes 9,380 sentences (average of 188 per document), annotated with seven rhetorical role labels.

(ii) Legal-Eval (Kalamkar et al., 2022) comprises judgments from the Indian Supreme Court. It contains 214 documents, with a total of 31,865 sentences (average of 115 sentences per document). Each sentence is annotated with 13 rhetorical role labels.

(iii) SCOTUS (Lavissière and Bonnard, 2024) includes 180 judgments from the Supreme Court of the United States. It contains a total of 22,600 sentences, with an average of 130 sentences per document, annotated with 13 rhetorical roles.

B Model Overview for Context Analysis

The model consists of four key components:

- **Word Embedding:** The target sentence and its retrieved context are encoded using BERT (Devlin et al., 2019), generating word-level embeddings.
- **Sentence Encoding:** A Bi-LSTM (Hochreiter, 1997) processes these embeddings, followed by attention-based pooling to obtain sentence representations.
- **Context Enrichment:** This layer models inter-sentence relationships to refine contextualized embeddings.
- **Output Layer:** A linear transformation maps the target sentence representation to logits, with labels predicted via softmax[7].

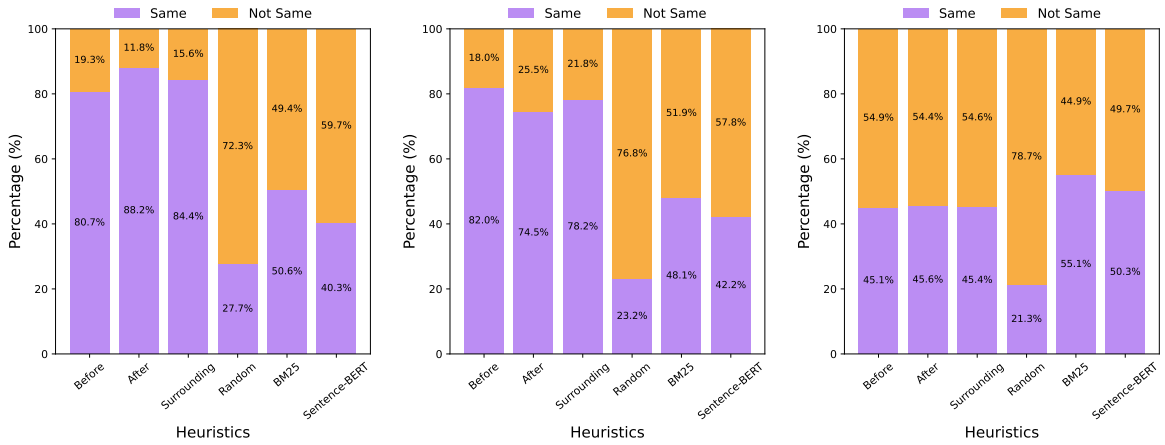
C Additional Results

We report additional results with enriching PLMs: RoBERTa (Liu et al., 2019), LegalBERT (Chalkidis et al., 2020), and Longformer (Beltagy et al., 2020) in Table 2.

[7]We optimize for the target sentence, eliminating the CRF layer, as supported by the ablation study in Jin and Szolovits (2018).

Model	Seq	DeepRhole	Legal Eval	SCOTUS
Roberta-base (baseline)	512	52.63	72.43	76.28
+ Before		**68.29**†	78.3†	**81.75**†
+ After		60.3†	**80.12**†	81.43†
+ Surrounding		63.86†	78.40†	80.10†
+ Random		50.04	72.35	75.79
+ BM25		53.54	72.79	77.78‡
+ Sentence-BERT		53.33	73.25‡	77.84‡
Legal-BERT (baseline)	512	54.06	69.43	76.85
+ Before		**69.10**†	79.65†	81.40†
+ After		63.19†	**80.99**†	**82.81**†
+ Surrounding		67.15†	78.55†	78.72
+ Random		50.32	68.55	76.56
+ BM25		54.59	70.77‡	77.06
+ Sentence-BERT		56.30	70.55	77.47
Longformer (baseline)	4096	53.83	72.57	76.26
+ Before		**67.62**†	79.89†	**81.58**†
+ After		61.16†	**80.09**†	81.09†
+ Surrounding		64.83†	73.09†	81.35†
+ Random		52.55	72.54	75.78
+ BM25		54.82	73.22	77.44†
+ Sentence-BERT		54.3	77.95‡	77.47‡

Table 2: Performance of PLMs using the best configuration observed in context analysis for $k \leq 6$ for each heuristic. Bold values represent the best improvement over the baseline (w/o context), while underlined values indicate the second-best. Markers † and ‡ denote statistical significance over the baseline at $p = 0.05$ and $p = 0.01$, respectively.



(a) DeepRhole (b) Legal Eval (c) SCOTUS

Figure 3: Analysis of retrieved sentences for each heuristic to determine the percentage of context sentences sharing the same label as the target sentence.

Target Sentence: *"This case focuses upon the requirement of 'fair presentation.'"*

Heuristic	Extracted Sentence
Before	*"O'Sullivan v. Boerckel, 526 U.S. 838, 845 (1999)."*
After	*"Michael Reese, the respondent, appealed his state-court kidnapping and attempted sodomy convictions and sentences through Oregon's state court system."*
Surrounding	*"O'Sullivan v. Boerckel, 526 U.S. 838, 845 (1999)."* *"Michael Reese, the respondent, appealed his state-court kidnapping and attempted sodomy convictions and sentences through Oregon's state court system."*
Random	*"In such instances, the nature of the issue may matter more than does the legal validity of the lower court decision."*
BM25	*"For another thing, the opinion-reading requirement would impose a serious burden upon judges of state appellate courts, particularly those with discretionary review powers."*
Sentence-BERT	*"The petition provides no citation of any case that might have alerted the court to the alleged federal nature of the claim."*

Table 3: Examples of sentences extracted using different heuristics from the SCOTUS dataset.