Faithfulness and Content Selection in Long-Input Multi-Document Summarisation of U.S. Civil Rights Litigation

Anonymous ACL submission

Abstract

Automatic summarisation of legal cases would reduce the burden on legal professionals and increase the accessibility of the law. However, the abstractive methods which dominate recent research are prone to hallucination. Despite the fact that this is a barrier to practical use, preventing hallucination is currently an under-studied area in the legal domain. We conduct the first study at the intersection of legal, multi-document, and faithful summarisation. In particular, by introducing a BERT-based content selection mechanism, we achieve an improvement of 0.2614 in the probability of a generated summary being entailed by its source text compared to a naïve content selection baseline, and observe qualitative improvements. Further, we demonstrate possible improvements of 5.56 ROUGE-1 F1, 5.46 ROUGE-2 F1, 2.7 ROUGE-L F1, and 2.15 BERTScore over the state-of-the-art if a perfectly predictive classifier was used, demonstrating the importance of content selection for summary faithfulness and quality for long-input legal abstractive summarisation.

1 Introduction

In common law jurisdictions, judicial decisions are informed by past cases, making identifying relevant precedent cases crucial (Zhong et al., 2020; Shukla et al., 2022; Askari et al., 2021; Althammer et al., 2021). However, the increasing number of precedent cases, each typically hundreds of pages long (Chalkidis et al., 2022), burdens legal professionals (Mumcuoğlu et al., 2021). While popular legal retrieval systems offer case summaries, these are costly and time-consuming to produce manually; automatic summarisation of legal cases using natural language processing tools would significantly benefit legal professionals, and increase the accessibility of the law.

However, abstractive summarisation methods are prone to hallucination - summaries may contain information which is unrelated or unfaithful to the source text (Feijo and Moreira, 2023). This is a major barrier to practical applicability (Huang et al., 2023; Wang et al., 2020; Fischer et al., 2022; Narayan et al., 2022a; Curran et al., 2023; Norkute et al., 2021), especially in the high-stakes domain of law (Feijo and Moreira, 2019, 2023); a Lexis-Nexis (2024) report found that 57% of respondents were concerned about hallucination. Despite this, the hallucination problem is understudied in relation to legal data. Additionally, the fact that the length of legal texts frequently exceeds transformer-based models' input token limit (Chalkidis et al., 2022) presents a challenging scenario.

Our work confronts these challenges by addressing the following research questions: **RQ1**: Can we improve the quality and faithfulness of abstractive summarisation results by providing a better representation of the source data to the summarisation model - namely, by using a BERT-based content selector trained on OREO labels to identify salient information? **RQ2**: Do transformer based models pretrained in the legal domain further improve results for legal multi document abstractive summarisation?

We contribute to the growing literature on faithfulness in abstractive summarisation, legal summarisation, and multi document summarisation by being the first work at this intersection. Specifically, we: (i) demonstrate that our content selection strategy improves summary faithfulness, through qualitative analysis and an improvement of 0.2614 in the probability of a generated summary being entailed by its source text compared to a naïve content selection baseline; and, (ii) demonstrate possible gains of 5.56 ROUGE-1 F1, 5.46 ROUGE-2 F1, 2.70 ROUGE-L F1, and 2.15 BERTScore if a perfectly predictive classifier was used for content selection in our methodology.

2 Related Work

2.1 Summarisation

Automatic summarisation methods aim to condense input text into a fluent shorter text retaining the key information (Feijo and Moreira, 2023; Kornilova and Eidelman, 2019; Bajaj et al., 2021). Extractive methods involve selecting and assembling key information from the source text (Jain et al., 2023), while recent abstractive summarisation methods, which are increasingly based on transformer architectures, generate summaries from scratch, conditioned on the source text.

The majority of legal summarisation research focuses on extractive summarisation, which is not our focus. Abstractive summarisation has been shown to significantly outperform extractive methods (Feijo and Moreira, 2019; Bhattacharya et al., 2019; Klaus et al., 2022), especially as transformer-based models pretrained on legal corpora (Shukla et al., 2022; Mullick et al., 2022; Chalkidis et al., 2020; Niklaus and Giofre, 2023; Zheng et al., 2021) have now been publicly released. Legal abstractive summarisation methodologies have investigated chunking (Shukla et al., 2022; Moro and Ragazzi, 2022), extractive summarisation (Bajaj et al., 2021), multitask learning (Elnaggar et al., 2018), argument roles (Xu et al., 2021; Elaraby and Litman, 2022), and prompt engineering (Pont et al., 2023). Although promising experimental results exist, the literature on legal abstractive summarisation is still relatively small, with multi-document summarisation being particularly understudied.

2.2 Faithfulness and Hallucination

Abstractive summarisation can lead to more natural summaries, but it may also introduce content unsupported by the source text, known as 'hallucination' (Huang et al., 2023; Nan et al., 2021b; Ji et al., 2023). Faithfulness refers to the consistency of the generated text with the input text (Sridhar and Visser, 2022), so reducing hallucination corresponds to increasing faithfulness (Ji et al., 2023).

Only one existing work (Feijo and Moreira, 2023) attempts to tackle the problem of hallucination for legal domain summarisation. Feijo and Moreira (2023) propose the LegalSumm method where summaries are generated for multiple distinct chunks of the source text, and a textual entailment model scores chunk-summary pairs to select the most faithful summary. However, this approach has limitations, such as the training examples to assess faithfulness not being reflective of real hallucination patterns, and the fact that the final summary derived from only one chunk may not include all salient information. Further, this method is not suitable for all judicial documents due to its use of specific case structure in the chunking process.

Various techniques to control hallucination have been proposed in the general domain, including filtering training examples (Matsumaru et al., 2020; Chaudhury et al., 2022), maximising faithfulness metrics during training (Nan et al., 2021b), modifying beam search (Zhao et al., 2020; Sridhar and Visser, 2022; Chaudhury et al., 2022; King et al., 2022), post-generation fact correction (Huang et al., 2023; Ji et al., 2023), and including additional information to guide generation (Dong et al., 2022; Cao et al., 2018; Narayan et al., 2022a, 2021).

3 Dataset

This study uses the Multi-LexSum dataset, the first dataset for legal multi-document summarisation (Shen et al., 2022). Multi-LexSum contains 9,280 expert-written summaries in accessible language pertaining to 4,539 U.S. civil rights lawsuits between 1950 and 2021, obtained from the Civil Rights Litigation Clearinghouse (CRLC). Case law was chosen for its practical application and volume. The documents to be summarised for each case include complaints, motions, and settlement agreements. Each of a case's documents can be over 100 pages, with a single case potentially involving hundreds of documents. A mean of 99378.2 words (10.3 documents) must be summarised per case, giving a very high compression ratio of 840.7.

Multi-LexSum contains multiple levels of summary granularity (examples in Appendix A); we focus on short summaries (mean 130 words), as long summaries (mean 646.5 words) frequently exceed the maximum decoder token length (1024) for the transformer model we use (PEGASUS). The short summaries cover the background, involved parties, and the case's outcome in a single paragraph. Writing summaries for standard cases takes 1-4 hours, while complex cases can take over 10 hours for an experienced lawyer.

Multi-LexSum is a relatively understudied dataset. Shen et al. (2022) conduct a preliminary study using Multi-LexSum using off-the-shelf models. While their results indicate that longer input lengths improve model performance, this is likely due to the content selection method used to handle maximum token length resulting in salient information not being included in the model input, which may also have led to hallucination by weakening the coupling between input-summary pairs during training. Human evaluation suggested that an alternative content selection strategy could thus enhance model performance and reduce hallucination.

3.1 Preprocessing

We applied preprocessing steps to the noisy Multi-LexSum data initially extracted using OCR. We detail this process and provide an annotated example before and after cleaning in Appendix B. The cleaning process allows the source text to be correctly segmented into sentences and paragraphs, which is vital in our methodology. As data filtering has been shown to minimize hallucinations (Nan et al., 2021a; Ji et al., 2023; Dong et al., 2022; Chaudhury et al., 2022; Narayan et al., 2021), we removed cases where less than 75% of legally salient entities in the summary occur in the source text. Additionally, to augment the dataset, we integrate long summaries (Shen et al., 2022) which are under 671 words, the maximum length for the short summary subset. Table 1 shows the dataset splits after cleaning, filtering, and augmentation.

	Complete Dataset	Short Summaries (Original)	Short Summaries (Preprocessed)
Train	4,539	3,138	3,436
Val.	3,177 (70%)	2,210 (70%)	2,508 (73%)
Test	454 (10%)	312 (10%)	312 (9%)
Total	908 (20%)	616 (20%)	616 (18%)

Table 1: Size of dataset splits after preprocessing.

4 Overview

The chosen task of abstractive summarisation involves generating a short summary S_i of the a set of N documents denoted as $D_i = \{D_{i_1}, D_{i_2}, \ldots, D_{i_N}\}$ belonging to the same case. We concatenate the documents D_i for each case in chronological order; the dates of each document were scraped from the CRLC website as these were not generally extractable from the text.

5 Models

We use PEGASUS (Zhang et al., 2020a), a state-of-the-art sequence-to-sequence transformer encoder-decoder model as our backbone abstractive summarisation model. Jointly with the Masked Language Modeling objective, PEGASUS has a pre-training objective designed specifically for abstractive summarisation - Gap Sentence Generation. Key sentences, selected based on ROUGE-F1, are masked from the input text during training, and the model must reproduce them; these key sentences are similar to a summary (Zhang et al., 2020a).

PEGASUS has a legal-pretrained variant, Legal-PEGASUS[1], trained on U.S. case law. Pretraining on legal data has been shown to increase performance on legal NLP tasks (Zhong et al., 2020; Shukla et al., 2022; Niklaus and Giofre, 2023). Shen et al. (2022) report results on PEGASUS and LED-16384 (a sparse attention transformer able to handle input lengths of up to 16,348 tokens (Beltagy et al., 2020)) that we use as baselines.

6 Content Selection

The self-attention mechanism in transformers limits the input token length to 1024 for PEGASUS (Zhang et al., 2020a). While sparse attention transformers (Beltagy et al., 2020; Guo et al., 2022; Zaheer et al., 2020) can handle longer input sequences and have shown promising performance in general-domain summarisation (Chalkidis et al., 2022; Niklaus and Giofre, 2023), in legal and multi-document cases, the input text often exceeds even these limits; in Multi-LexSum, the average source text length for a case is 83,340 tokens, with a maximum of 4,423,683 tokens. Thus, a content selection strategy to ensure salient information is included within this input token limit is essential; if the input to the summarisation model does not contain the relevant information, summary quality is reduced and hallucination is encouraged, as the input-summary pairs are not tightly coupled.

Previous approaches to handling the input token limit include segmenting the source text in chunks and then concatenating summaries. However, this introduces a number of issues: it is non-trivial to extract the corresponding sentences from the reference summary for each chunk, not all chunks may be equally informative, independent chunk processing may lead to redundancy in the final summary, and for long input texts, summarising every chunk is computationally expensive (Shukla et al., 2022; Moro and Ragazzi, 2022). Similarly, multi-stage frameworks (Zhang et al., 2022), which iteratively use the concatenated summary as the input to another phase of chunking and abstractive summarisa-

[1] https://huggingface.co/nsi319/legal-pegasus

tion, significantly increase computational complexity and introduce opportunities for hallucination.

To address these issues, we propose a mixed-model approach. We first identify salient information from the source text, and then use this information as input to our backbone model. Importantly in the legal domain, this approach better mirrors the human summarisation process, and hence may contribute to user trust. There is evidence that a human summarising long input text would highlight the important information and then paraphrase this information to form a summary (Bajaj et al., 2021; Norkute et al., 2021; Liu et al., 2018; Jing and McKeown, 1999), and a study on legal text summarisation demonstrates participants' increased trust in systems for which they understand the summary's creation process and feel that this process is similar to their own (Norkute et al., 2021; Danilevsky et al., 2020; Adadi and Berrada, 2018).

6.1 Oracle Extracts for Gold Labels

We adopt a ranking-based approach to select salient information by training a BERT-based salience classifier to extract relevant sentences from the source text, using the state-of-the-art OREO[2] method to obtain gold standard training labels. This enables us to create a list of all source text sentences ranked by the classifier's confidence that the sentence contains salient information. During inference, the top-ranked sentences are utilized to construct the input to the PEGASUS model (when finetuning PEGASUS, we instead use the gold standard sentences from OREO as the model input).

To obtain 'gold standard' data regarding which sentences of the source text contain salient information for summarisation purposes, we must convert the gold-standard abstractive summaries to their extractive equivalent. As annotations by legal professionals would be prohibitively costly and time-consuming, we use an automatic labeling approach.

Various approaches have been proposed to create oracle extracts, among which greedily maximising the ROUGE overlap with the gold-standard summary is most common (Xu and Lapata, 2022; Bhattacharya et al., 2021; Klaus et al., 2022). However, oracles constructed in this way do not always lead to high-performing summaries (Xu and Lapata, 2022) - indeed, a recent study on legal extractive summarisation (Klaus et al., 2022) suggests that 'alternative methods to create oracle extractive summaries' should be considered. Furthermore, this greedy approach considers only a *single* oracle summary, Y^*, but there can be *multiple* valid oracle summaries for the same source text; systems trained on greedy oracles are optimised by maximising the probability at Y^* and assigning zero probability to all other hypotheses, regardless of quality. For this reason, we use the OREO algorithm to create oracles, which incorporates the idea of learning from *multiple* oracle summary hypotheses. Xu and Lapata (2022) showed that OREO led to superior performance compared to the common greedy approach, and that OREO can better guide the learning and inference of an abstractive summarisation system. Further details and hyperparameter details are provided in Appendix C; here we note that OREO is fundamentally ROUGE based.

[2] https://github.com/yumoxu/oreo

6.2 Sentence Salience Classification

Using the 'oracle' sentences output by OREO, we train a classifier to determine the summary-worthiness of sentences (i.e. the binary label assigned by OREO). Conceptually, as obtaining OREO labels requires a case to already have a gold-standard summary, training a classifier to *predict* which sentences of a legal case's source text are summary-worthy (by training the classifier at a sentence level on OREO labels) allows us to carry out content selection on unseen cases.

We use a legal oriented pre-trained model, CaseLawBERT (Zheng et al., 2021), which provides the best domain match, and achieve an ROC-AUC score of 0.884 (curve in Appendix D). Due to class imbalances (Table 2), we conduct random downsampling, resulting in 68,592 training examples. However, addressing this imbalance in a more sophisticated manner may lead to improved results.

	All Instances	Positive Instances	Negative Instances
Train	6,230,772	34,296 (0.55%)	6,196,476
Val.	1,122,744	4,355 (0.39%)	1,118,389
Test	1,672,233	8,021 (0.48%)	1,664,212

Table 2: Number of instances of each class (binary, assigned by OREO) for sentence salience classification (before downsampling). Positive instances are considered summary-worthy.

6.3 Input Construction

To construct the PEGASUS inputs from the ranked list of sentences, we compare several strategies, adding tokens until the limit of 1024 is reached:

- Sentences - we add the top scoring sentences with non-zero scores, as in Xu and Lapata (2022).

- Windows - we add the preceding and following sentence for each selected sentence. This provides context, but may lead to irrelevant information being included.

- Paragraphs - we add the whole paragraph the sentence is contained within.

In all cases, we concatenate the extracted information in order of appearance in the temporally ordered source documents. We consider two variants: BERT (ranked list of sentences is obtained from training the BERT classifier on OREO labels), and OREO (ranked list of sentences is obtained directly from OREO, to investigate the *potential* gains our content selection strategy could produce if the classifier was perfectly predictive).

We also consider three baseline methods:

- First-1024 - we take the first 1024 tokens of the temporal concatenation of all the case's source documents.

- First-K - like in the original MultiLex-sum paper (Shen et al., 2022), for a case with *D* documents, we take the first *1024/D* tokens of each. Unlike Shen et al. (2022), the dataset has been cleaned and temporally ordered.

- TextRank - a general-domain unsupervised extractive summarisation method, frequently used as a content selection baseline (Liu et al., 2018; Bajaj et al., 2021; Klaus et al., 2022).

6.4 Content Selection Preliminary Results

As a preliminary experiment, we investigate the ROUGE recall between the *extracts* produced (which will be used as input to PEGASUS) and the corresponding gold standard summary for the test set (as we have already performed the expensive inference process for the BERT classifier on test set data). We use recall as we wish to consider if the salient information has been selected, not the specificity of salient information. Results are presented in Table 3.

BERT-Sentences and BERT-Windows outperform the naive First-1024 and First-K baselines, with TextRank also performing well. However, the BERT-Paragraphs method performs poorly, likely due to including too much context for each selected sentence and thus being able to include fewer highly-ranked sentences. We also compare the three BERT-based strategies to their OREO counterparts. OREO-Windows performed best overall in terms of ROUGE-1 and ROUGE-2 recall. All OREO strategies outperformed their BERT-based counterparts in terms of ROUGE-2, although BERT-Sentences outperformed OREO-Sentences in terms of ROUGE-1. We also noted that OREO extracted significantly fewer tokens than BERT in the sentence case, suggesting that an input token length of 1024 tokens is sufficient.

	ROUGE-1	ROUGE-2	ROUGE-L
First-1024	67.51	24.35	41.50
First-K	57.36	19.25	35.94
TextRank	70.28	23.47	43.93
BERT-Sentences	**76.61**	**32.61**	**46.15**
BERT-Windows	73.88	28.00	41.95
BERT-Paragraphs	58.30	19.99	32.66
OREO-Sentences	68.07	32.73	35.43
OREO-Windows	**79.43**	**37.13**	**45.25**
OREO-Paragraphs	73.83	33.24	41.59

Table 3: Mean ROUGE recall scores against corresponding gold standard summary for each strategy tested.

7 Experimental Setup

Overall, we vary two dimensions in our experiments, corresponding to our research questions: input representation (RQ1, Section 6), and domain match (RQ2, Section 5).

For comparison to the PEGASUS results reported for Multi-LexSum in Shen et al. (2022), we use the same hyperparameters values where provided: we train for 6 epochs with a learning rate of 5e-5, and for inference we use beam search with 5 beams and n-gram repetition blocks for n>3. For additional hyperparameters, we trained the models with a batch size of 4, 64 gradient accumulation steps, gradient checkpointing enabled, and a weight decay of 0.01. For our models at inference, we used a minimum of 24 tokens and maximum of 960 tokens for experimental settings with no entity chain, and a minimum of 34 tokens and maximum of 1154 for experimental settings including some form of entity chain, as these were the boundaries observed for our gold-standard data. We also added a length penalty of 2.0 to encourage the generation of long sequences, as Shen et al. (2022) observed that PEGASUS undergenerated the number of words when producing short summaries for Multi-LexSum.

8 Evaluation

We evaluate the quality of the produced summaries using standard ROUGE-1, ROUGE-2, and ROUGE-L scores. We also report BERTScore (Zhang et al., 2020b) to capture semantic similarity without relying solely on lexical overlap, as ROUGE fails to capture deeper semantic similarity (Shukla et al., 2022; Bhattacharya et al., 2019; Kanapala et al., 2019; Jain et al., 2023; Zhong et al., 2019; Cohan and Goharian, 2016; Kikuchi et al., 2014; Feijo and Moreira, 2023). We used the DeBERTA model for comparison with previous work.

We evaluate faithfulness using textual entailment following Narayan et al. (2022b,a) and previous studies demonstrating a correlation between entailment scores and human judgements of faithfulness (Narayan et al., 2022b; Fischer et al., 2022; Sridhar and Visser, 2022; Kryscinski et al., 2020; Maynez et al., 2020; Honovich et al., 2022). We report the probability of a generated summary (PEGASUS output) being entailed by its source text (PEGASUS input) returned by a BART-large classifier finetuned on Multi-NLI (Fischer et al., 2022).

9 Results and Analysis

9.1 Input Representation

To first investigate which content selection approaches are promising, for the standard (not legal pretrained) variant of PEGASUS, Table 4 shows ROUGE and BERTScore F1 scores for our baselines, the three BERT-based strategies, the three additional OREO-based strategies, the PEGASUS and state-of-the-art results reported in Shen et al. (2022), and our reproduction of the PEGASUS results (needed due to incomplete knowledge of hyperparameters used in Shen et al. (2022)).

Among BERT-based strategies, BERT-Windows, our most effective method and improved **ROUGE-1 by 0.82** compared to the reported PEGASUS performance. However, we do not observe improvements with respect to other metrics.

On all metrics apart from ROUGE-2, BERT-Windows was the most effective of the six tested input strategies. This is likely due to the balance of the number of relevant sentences included and providing context for each sentence. First-1024, First-K, BERT-Sentences, and BERT-Windows all outperform our reproduction of Shen et al. (2022)'s with respect to ROUGE-1, as expected, and BERT-Sentences and BERT-Windows outperform the reproduction baseline with respect to ROUGE-2. TextRank fails to outperform this baseline, which is consistent with its poor performance as a content selector for abstractive summarisation in Bajaj et al. (2021). BERT-Paragraph also fails to outperform the baseline, likely due to including longer context for each sentence, which limits the relevant information that can be included.

None of our 6 proposed strategies outperform the reproduction baseline on ROUGE-L or BERTScore metrics. We expected a greater improvement from the First-K baseline over the reproduction baseline, which intuitively should improve results as its only difference to the content selection strategy in Shen et al. (2022) is the introduction of dataset cleaning and temporal ordering. We hypothesise that the dataset filtering process may have resulted in decreased ROUGE scores[3], consistent with Nan et al. (2021a) - although this is likely to contribute to increased faithfulness.

To investigate the potential of content selection strategies independently of the BERT classifier (i.e. if the BERT salience classifier was perfectly predictive of OREO labels), we also analyzed the model's performance using the oracles from OREO as inputs. The OREO strategies outperformed BERT counterparts, with OREO-Sentences surpassing the SOTA by up to **4.45 ROUGE-1**, **4.39 ROUGE-2**, **1.40 ROUGE-L**, and **0.27 BERTScore**.

As OREO-Sentences extracts typically consist of far fewer tokens (mean 264.15) than BERT-sentences (mean 1000.78), yet BERT-Sentences extracts have a greater ROUGE recall with the reference summary (see Section 6.4), this suggests that the specificity and saliency of the inputs provided to PEGASUS is key. Indeed, when measuring the mean of ROUGE-1 and ROUGE-2 precision between the OREO and BERT extracts used as input to PEGASUS with the gold summary, the OREO extracts display a greater precision (28.57 vs 7.67 for sentences, 10.68 vs 7.73 for windows, and 12.82 vs 12.15 for paragraphs). The increasing similarity in precision scores between OREO and BERT variants as the number of sentences for which information is included in the extracts decreases also suggests that the BERT classifier performs best for its high confidence outputs.

Overall, we establish that content selection does have the potential to improve summarisation out-

[3] This was *not* due to the augmentation process - we performed an ablation study without dataset augmentation for the Lead-K baseline and achieved poorer results: ROUGE-1 F1 42.56, ROUGE-2 F1 18.41, ROUGE-L F1 27.93.

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	Entailment
			Baselines		
PEGASUS (reproduced)	43.23	19.26	29.35	36.15	0.2937
PEGASUS	43.35	19.91	29.99	37.88	-
LED-16384 (SOTA)	46.54	22.08	31.91	40.00	-
			PEGASUS		
First-1024	43.39	18.96	28.42	34.47	-
First-K	43.24	18.96	28.40	34.94	-
TextRank	42.36	17.23	27.31	33.45	-
BERT-Sentences	43.61	19.33	27.58	34.52	0.5134
BERT-Windows	44.17	19.28	28.53	35.62	0.5551
BERT-Paragraphs	40.14	16.28	25.95	31.39	-
OREO-Sentences	50.99	26.47	33.31	40.27	0.4915
OREO-Windows	47.97	23.28	31.55	38.92	0.5457
OREO-Paragraphs	47.15	22.42	30.83	37.84	-
			Legal-PEGASUS		
BERT-Sentences	42.77	19.08	27.25	34.81	0.4954
BERT-Windows	44.34	19.55	28.91	36.35	0.5551
OREO-Sentences	52.10	27.54	34.61	42.15	0.4680
OREO-Windows	48.41	23.72	31.91	39.44	0.5469

Table 4: The upper part shows results for PEGASUS summaries, the lower part for Legal-PEGASUS summaries. We report Mean ROUGE and BERTScore F1 scores with respect to the corresponding reference summary. The last column shows entailment scores (not calculated for all experimental setups due to limited compute resources). We highlight the best scores for OREO and BERT in red and blue respectively, for PEGASUS and Legal-PEGASUS.

puts, but that the salience classifier performance limits these improvements in practice.

9.2 Domain-Specific Pretraining

As the sentence and window-based strategies offer the most promising results, we only report Legal-PEGASUS results for these strategies. The lower part of Table 4 shows results for Legal-PEGASUS. With legal pretraining, we observe improvements in BERTScore and ROUGE-1 for all input settings. Our best results for the complete pipeline are given by BERT-Windows. However, these results still only outperform the PEGASUS results reported in Shen et al. (2022) with respect to ROUGE-1, by **0.99 F1**. In contrast, OREO-Sentences further outperforms the state of the art, achieving an improvement of **5.56 ROUGE-1 F1**, **5.46 ROUGE-2 F1**, **2.7 ROUGE-L F1**, and **2.15 BERTScore**. Overall, we observed greater improvements for better content selection strategies. Our results again indicate the importance of content selection, and the importance of the domain match at pretraining.

9.3 Faithfulness

Entailment scores are reported in Table 4. While we do not have entailment scores for the exact PEGASUS setup in Shen et al. (2022) as we do not have access to the original model outputs, and we acknowledge that our reproduction leads to slightly different results, it is evident that all our experimental setups vastly improve the probability of the source text entailing the summary text in comparison to this reproduction baseline (mean entailment probability 0.2937); our BERT-Windows content selection strategy improves entailment probability by 0.2614 for both PEGASUS and Legal-PEGASUS. This suggests that content selection is effective in improving summary faithfulness.

Interestingly, Legal-PEGASUS led to reduced entailment scores compared to vanilla PEGASUS. Overall, BERT-based methods consistently exhibited higher faithfulness than OREO-based counterparts, and window-based methods showed higher faithfulness than sentence-based methods. Our findings align with the literature in that ROUGE does not correlate with faithfulness; although OREO-Sentences receives the worst entailment scores, this method performs best on ROUGE and BERTScore.

9.4 Qualitative Analysis

Although human expert evaluation of the summaries is infeasible, to better understand our models' behaviour and failure modes, we manually analysed generated summaries for a sample of 10 cases across experimental settings (example in Appendix G). In general, the outputs of the reproduction of the PEGASUS method in Shen et al. (2022) were comparatively good at reproducing the correct date when the case began, as this is frequently mentioned at the start of the document. Background information for the case (often at the start of the initial document) are also reflected fairly reliably. However, the summaries often hallucinate the law which is alleged to be violated, which is extremely vital, and struggle to accurately represent the case's procedure. This is likely as this information is not included in the input text captured using the naïve content selection strategy.

In contrast, our models produce longer summaries which better match the reference summaries in content (not reflected by the ROUGE results). We observed limited variation across input strategies, including OREO-based strategies. In general, our models perform well for the background and laws involved in the case, but performance often declines for a case's procedural actions, with key information being missed, and the reasoning for decisions failing to be provided. This may be because these aspects follow a less standard format, so are less easily identified by the BERT classifier (Zhong et al., 2019). While our models contain less hallucinatory content than our reproduction of Shen et al. (2022)'s approach, two common hallucination scenarios remain. Firstly, dates and monetary amounts which occur in the source text are often contained in the summary in the incorrect context; such intrinsic hallucination is non-trivial to combat. The second scenario stems from issues relating to case understanding. A notable subtype of this is the inclusion of information from *cited cases* as if it pertains to the main case under discussion; this may be because discussions of cited cases often include a high density of common legal keywords. As both BERT and OREO methods make this mistake, this suggests that selecting relevant sentences may be more suited to human annotation than automatic overlap-based methods; while full-scale human annotation would be infeasible, semi-supervised approaches, which have been applied with success in other areas of legal AI (Branting et al., 2019), may be promising. At the BERT classifier stage, including context for the sentence under consideration may help to distinguish between information relating to main or cited cases.

Another issue is our models' tendency to include large extractive fragments from the source text, evidenced by artefacts (such as numerals) remaining despite the cleaning process being reproduced. This limits the readability of the summaries in some cases by replicating complex legal terminology and syntax from the source text. This high degree of extractivity may be due to limited text occurring in the model input for each point, and that these fragments are not well-flowing text, which models such as PEGASUS are trained on.

10 Discussion and Conclusion

We conduct the first study at the intersection of legal, multi-document, and faithful summarisation, investigating the impact of content selection and legal pretraining on the abstractive summarisation of U.S. civil rights litigation, with PEGASUS as our backbone model. Our full test-time pipeline outperforms the PEGASUS results in Shen et al. (2022) by 0.99 ROUGE-1 F1. We show that using oracle extracts vastly outperforms the state-of-the-art, with legal pretraining further boosting results: we achieve an improvement of 5.56 ROUGE-1 F1, 5.46 ROUGE-2 F1, 2.7 ROUGE-L F1, and 2.15 BERTScore. Our content selection strategy also leads to an improvement of 0.2614 in the probability of a generated summary being entailed by its source text, compared to a naïve content selection baseline. Overall, we provide evidence that content selection has the ability to improve summary faithfulness and quality. However, the generated summaries can still contain hallucinations and omit key information. Several issues, such as the quality of the content selection method and addressing specific hallucination scenarios, remain to be addressed for such automatic summarisation to see real-world adoption.

Our study's limitations and error cases suggest several future research areas. Further research into content selection is promising - investigating methods of content selection not fundamentally based on ROUGE, such as using human salience annotations in a semi-supervised framework, could be fruitful. More generally, the application of our generated summaries as the input to other legal NLP tasks could be studied. Also, future work could conduct similar investigations on different legal domains and jurisdictions, or using different backbone models; for example, it would be interesting to observe the effect of content selection on models able to handle longer inputs, such as LED, or GPT-based models.

11 Limitations

Our project is limited by the resources available. Firstly, while presenting a realistic summarisation scenario, the OCR process used to construct the Multi-LexSum dataset from the original court documents introduces noise, which despite dataset cleaning, can adversely affect both summarisation outputs and the metrics used to evaluate these outputs. The availability of computational resources also limits the range of experiments that can be conducted and the hyperparameter settings used. Perhaps most importantly, the lack of a thorough human evaluation of our models' outputs by domain experts limits our interpretation of our findings, as metrics such as ROUGE and entailment are only *proxies* for summary quality and faithfulness. This factor is especially important in the legal domain, where a lack of correlation between automatic metrics and human expert judgements has previously been demonstrated (Shukla et al., 2022; Bhattacharya et al., 2019), and as the utility of automatic metrics to judge faithfulness remains a topic of research debate.

Our work also has limitations pertaining to the intended use case of legal summarisation - namely, by legal professionals or ordinary civilians without resources or expertise in machine learning. The powerful GPUs required for finetuning and performing inference for the transformer models used throughout our pipeline are unlikely to be available in non-academic environments. Furthermore, our methodology's performance on other datasets (for example - for other legal areas, jurisdictions, or languages) has not yet been tested.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160. Conference Name: IEEE Access.

Sophia Althammer, Arian Askari, Suzan Verberne, and Allan Hanbury. 2021. DoSSIER@COLIEE 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval. ArXiv:2108.03937 [cs].

A. A. Askari, S. V. Verberne, O. Alonso, S. Marchesin, M. Najork, and G. Silvello. 2021. Combining lexical and neural retrieval with longformer-based summarization for effective case law retrieva. In *Proceedings of the second international conference on design of experimental search & information REtrieval systems*, pages 162–170. CEUR.

Ahsaas Bajaj, Pavitra Dangati, Kalpesh Krishna, Pradhiksha Ashok Kumar, Rheeya Uppaal, Bradford Windsor, Eliot Brenner, Dominic Dotterrer, Rajarshi Das, and Andrew McCallum. 2021. Long Document Summarization in a Low Resource Setting using Pretrained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 71–80, Online. Association for Computational Linguistics.

Claire Barale, Michael Rovatsos, and Nehal Bhuta. 2023. Automated Refugee Case Analysis: A NLP Pipeline for Supporting Legal Practitioners. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2992–3005, Toronto, Canada. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv:2004.05150 [cs].

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 413–428, Cham. Springer International Publishing.

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 the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pages 22–31, New York, NY, USA. Association for Computing Machinery.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Michael James Bommarito, Daniel Martin Katz, and Eric Detterman. 2018. LexNLP: Natural Language Processing and Information Extraction For Legal and Regulatory Texts.

K. Branting, B. Weiss, B. Brown, C. Pfeifer, A. Chakraborty, L. Ferro, M. Pfaff, and A. Yeh. 2019. Semi-Supervised Methods for Explainable Legal Prediction. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, pages 22–31, New York, NY, USA. Association for Computing Machinery.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: fact-aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium* *on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pages 4784–4791, New Orleans, Louisiana, USA. AAAI Press.

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.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, and Alexander Gray. 2022. X-FACTOR: A Cross-metric Evaluation of Factual Correctness in Abstractive Summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7100–7110, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2016. Revisiting Summarization Evaluation for Scientific Articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813. European Language Resources Association (ELRA).

Shawn Curran, Sam Lansley, and Oliver Bethell. 2023. Hallucination is the last thing you need. ArXiv:2306.11520 [cs].

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the Document or to the World? Mitigating Hallucinations via Entity-Linked Knowledge in Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ahmed Elnaggar, Christoph Gebendorfer, Ingo Glaser, and Florian Matthes. 2018. Multi-Task Deep Learning for Legal Document Translation, Summarization and Multi-Label Classification. In *Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference*, AICCC '18, pages 9–15, New York, NY, USA. Association for Computing Machinery.

Diego Feijo and Viviane Moreira. 2019. Summarizing Legal Rulings: Comparative Experiments. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 313–322, Varna, Bulgaria. INCOMA Ltd.

Diego de Vargas Feijo and Viviane P. Moreira. 2023. Improving abstractive summarization of legal rulings through textual entailment. *Artificial Intelligence and Law*, 31(1):91–113.

Tim Fischer, Steffen Remus, and Chris Biemann. 2022. Measuring Faithfulness of Abstractive Summaries. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 63–73, Potsdam, Germany. KONVENS 2022 Organizers.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient Text-To-Text Transformer for Long Sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating Factual Consistency Evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2023. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey. ArXiv:2104.14839 [cs].

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2023. Bayesian Optimization based Score Fusion of Linguistic Approaches for Improving Legal Document Summarization. *Knowledge-Based Systems*, 264:110336.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):248:1–248:38.

Hongyan Jing and Kathleen R. McKeown. 1999. The decomposition of human-written summary sentences. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 129–136, New York, NY, USA. Association for Computing Machinery.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51(3):371–402.

Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2014. Single Document Summarization based on Nested Tree Structure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 315–320, Baltimore, Maryland. Association for Computational Linguistics.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. 2022. Don't Say What You Don't Know: Improving the Consistency of Abstractive Summarization by Constraining Beam Search. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 555–571, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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*, SIGIR '22, pages 2426–2430, New York, NY, USA. Association for Computing Machinery.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A Corpus for Automatic Summarization of US Legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

LexisNexis. 2024. Lawyers cross into the new era of generative AI. Technical report, LexisNexis.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. ArXiv:1801.10198 [cs].

Jayawant N. Mandrekar. 2010. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving Truthfulness of Headline Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Gianluca Moro and Luca Ragazzi. 2022. Semantic Self-Segmentation for Abstractive Summarization of Long Documents in Low-Resource Regimes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11085–11093. Number: 10.

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

Emre Mumcuoğlu, Ceyhun E. Öztürk, Haldun M. Ozaktas, and Aykut Koç. 2021. Natural language processing in law: Prediction of outcomes in the higher courts of Turkey. *Information Processing and Management: an International Journal*, 58(5).

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021a. Entity-level Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021b. Improving Factual Consistency of Abstractive Summarization via Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6881–6894, Online. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Dipanjan Das, and Mirella Lapata. 2022a. Conditional Generation with a Question-Answering Blueprint. ArXiv:2207.00397 [cs].

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022b. A Well-Composed Text is Half Done! Composition Sampling for Diverse Conditional Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1339, Dublin, Ireland. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with Learned Entity Prompts for Abstractive Summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Joel Niklaus and Daniele Giofre. 2023. Can we Pretrain a SotA Legal Language Model on a Budget From Scratch? In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 158–182, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. 2021. Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Vedant Parikh, Vidit Mathur, Parth Mehta, Namita Mittal, and Prasenjit Majumder. 2021. LawSum: A weakly supervised approach for Indian Legal Document Summarization. ArXiv:2110.01188 [cs].

Thiago Dal Pont, Federico Galli, Andrea Loreggia, Giuseppe Pisano, Riccardo Rovatti, and Giovanni Sartor. 2023. Legal Summarisation through LLMs: The PRODIGIT Project. ArXiv:2308.04416 [cs].

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-world Summaries of Civil Rights Lawsuits at Multiple Granularities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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.

Arvind Krishna Sridhar and Erik Visser. 2022. Improved Beam Search for Hallucination Mitigation in Abstractive Summarization. ArXiv:2212.02712 [cs].

Kazem Taghva, Tom Nartker, Allen Condit, and Julie Borsack. Automatic Removal of "Garbage Strings" in OCR Text: An Implementation.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

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

Huihui Xu, Jaromir Savelka, and Kevin D. Ashley. 2021. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pages 250–254, New York, NY, USA. Association for Computing Machinery.

Yumo Xu and Mirella Lapata. 2022. Text Summarization with Oracle Expectation. ArXiv:2209.12714 [cs].

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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 11328–11339. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating Text Generation with BERT. ArXiv:1904.09675 [cs].

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing Quantity Hallucinations in Abstractive Summarization. In *Findings of the Association* for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pages 159–168, New York, NY, USA. Association for Computing Machinery.

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.

Linwu Zhong, Ziyi Zhong, Zinian Zhao, Siyuan Wang, Kevin D. Ashley, and Matthias Grabmair. 2019. Automatic Summarization of Legal Decisions using Iterative Masking of Predictive Sentences. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, pages 163–172, New York, NY, USA. Association for Computing Machinery.

A Summary Granularities in Multi-LexSum

Here we present examples of the three summary granularities in Multi-LexSum: long, short, and tiny.

Source Input Excerpt ... And, even if the agency had made an internal decision to maintain the status quo, the documents at issue would not lose their predecisional status because plaintiff has not shown that they have been "adopted, formally or informally, as the agency position on an issue or is used by the agency in its dealings with the public."1 Coastal States Gas Corp., 617 F.2d at 866; Sears, 421 U.S. at 161 ("[I]f an agency chooses expressly to adopt or incorporate by reference an intraagency memorandum previously covered by Exemption 5 in what would otherwise be a final opinion" that memorandum may not be withheld under Exemption 5). Plaintiff does not point to any public statements that OMB has made referencing, adopting, or incorporating the records or the subject matter at issue, nor has plaintiff provided the Court with any evidence that the records were informally adopted as the agency's position. Plaintiff references a statement made by Karen Battle, chief of the Census Bureau's Population Division, on January 26, 2018, where she explained that additional research and testing were necessary before the Census Bureau could proceed to implement a separate Middle Eastern or North African category. Pl.'s Cross-Mem. at 13. Plaintiff argues that "[t]o the extent that Ms. Battle's explanation about the need for more research, and indeed the entire underlying decision to maintain the status quo, is evidenced in the withheld documents, it has been adopted as the agency's policy." Id. But, this statement was made by a Census Bureau official, not an OMB official. And, in any event, the statement 1 Courts in this district have held that the plaintiff carries the burden to show that the agency has formally or informally adopted a record as policy. See, e.g., Heffernan v. Azar, 317 F. Supp. 3d 94, 122 (D.D.C. 2018), citing Sec. Fin. Life Ins. Co. v. U.S. Dep't of Treasury, No. 03-102, 2005 WL 839543, at *7 (D.D.C. Apr. 12, 2005). ...

Long Summary: On April 13, 2018, the Arab American Institute ("AAI") sued the Office of Management and Budget ("OMB") under the Freedom of Information Act ("FOIA"), 5 U.S.C. § 552, in the U.S. District Court for the District of Columbia. AAI alleged that OMB violated FOIA by failing to disclose requested records pertaining to OMB's decision not to include a combined race and ethnicity question or a Middle Eastern or North African (MENA) category on the 2020 Census. AAI asked the court to declare that OMB violated FOIA and to issue an injunction ordering the agency to release the requested records. This case was assigned to Judge Amy Berman Jackson One month later, on May 18, 2018, the court ordered OMB to file a dispositive motion or a status report setting a schedule for OMB's production of documents to AAI. OMB chose the latter, filing its first status report on June 15, 2018. Over the next two years, the parties filed several joint status reports detailing which documents OMB had disclosed to AAI and which documents were still outstanding or disputed. By May 13, 2020, OMB had reviewed approximately 2,000 potentially responsive documents, producing "a number" of them to AAI and withholding 161 of them, claiming they were FOIA exempt. AAI objected to the withholding of five of the allegedly exempt documents. OMB filed a motion for summary judgment on February 10, 2020, arguing that the five disputed documents were exempt under FOIA Exemption 5, which allows agencies to withhold "inter-agency or intra-agency memorandums or letters that would not be available by law to a party other than an agency in litigation with the agency," including "predecisional and deliberative" documents that reflect internal Executive Branch deliberations. AAI filed a cross-motion for summary judgment on March 12, 2020, arguing that OMB had not provided a sufficient basis for exempting the documents and that the exemption didn't apply because the documents were not "predecisional." On August 13, 2020, after conducting in camera review, the court granted OMB's motion for summary judgment and denied AAI's cross-motion, finding that the disputed documents were predecisional and exempt from FOIA. 2020 WL 4698098. As of December 25, 2020, AAI has not appealed the court's decision.

Short Summary: On April 13, 2018, the Arab American Institute sued the Office of Management and Budget under the Freedom of Information Act in the U.S. District Court for the District of Columbia. AAI alleged that OMB violated FOIA by failing to disclose requested records pertaining to OMB's decision not to include a combined race and ethnicity question or a Middle Eastern or North African (MENA) category on the 2020 Census. In May, the court ordered OMB to file a dispositive motion or a status report setting a schedule for OMB's production of documents to AAI. Over the next two years, the parties filed several joint status reports detailing which documents OMB had disclosed to AAI and which documents were still outstanding or disputed. OMB produced a number of documents to AAI but withheld some, claiming they were FOIA exempt. AAI objected to five claimed exemptions. The parties both filed motions for summary judgment. After conducting in camera review, on August 13, 2020, the court granted OMB's motion for summary judgment and denied AAI's cross-motion, finding that the disputed documents were predecisional and exempt from FOIA. As of December 25, 2020, AAI has not appealed the court's decision.

Tiny Summary: The Office of Management and Budget is forced to disclose documents requested by the Arab American Institute under the Freedom of Information Act. (D.D.C.)

B Document Cleaning

The use of OCR (as required in real-world scenarios) to obtain plain text data from PDF court documents (Shen et al., 2022) of variable legibility containing formatting such as headers, footnotes, citations, and tables results in the source text in the Multi-LexSum dataset containing errors and noise. Therefore, despite the underlying quality of the judicial documents, we first conducted dataset cleaning to allow for subsequent steps such as segmentation to be meaningfully applied, as in many cases we find 'junk' in the middle of paragraphs or sentences, and erroneous line breaks.

The overall cleaning pipeline for each source document is illustrated in Figure B. To define the rules for cleaning, we studied the text in the Multi-LexSum dataset and the corresponding original documents available on the CRLC website for cases in the validation set. For each newly implemented rule, we tested their validity on subsequent documents in the validation set, and ensured that previously considered documents were not adversely affected. This process continued until a stable set of rules was reached, which was then applied to all source documents.

- Removal of footers: We removed document footers containing irrelevant entities.
- Removal of headers: We only keep lines meeting at least one of the following conditions:
  - The line stripped of numerals only occurs once in the document - headers occur multiple times in the document, but may contain page numbers; thus, when stripped of numerals, this stripped line occurs multiple times.
  - The length of the stripped line is less than 20 characters - headers are long, we do not want to remove other information which may be repeated throughout the document, such as names, or terms such as 'v.' or 'and'.
  - The line does not contain any numerals or hyperlinks - headers usually contain one or both, and we do not want to remove useful information.
- Removal of dirty lines: Dirty lines include page numbers, hyperlinks, lines not containing alphabetical characters, timestamps, and 'junk' resulting from OCR. Timestamp lines were identified using the *dateutils* parser. To remove 'junk' lines resulting from the OCR process, we edited *garbage_detector*[4] (based on Taghva et al.), which identifies a line of text as 'garbage' if any one of several given conditions holds. We

[4] https://github.com/foodoh/rmgarbage

Figure 1: Summary of main stages of the cleaning pipeline.

removed two of the conditions originally provided, as these gave many false positives in the legal domain: uppercase between lowercase; two distinct punctuation marks in the same line. We kept the remaining three original conditions, relating to a string's ratio of alphanumeric characters to total characters, ratio of consonants to vowels, and if a punctuation mark repeats consecutively (this condition was edited to reflect the fact that while periods and brackets can legitimately repeat consecutively, punctuation marks such as commas, colons, semicolons, and dollars cannot). Finally, we added a condition to capture the fact that certain punctuation marks appearing between lower-case letters is indicative of junk text.

- Line breaks: This includes removing blank lines, removing newlines in the middle of sentences or paragraphs, and correctly ensuring a newline before each new legal paragraph. We kept existing line breaks only after colons (used to precede legal lists), after periods where the previous character was not a capital letter or 'v' (to avoid line breaks after abbreviations such as v. or U.S.), or if the whole line consisted of upper case letters (indicative of a section title). To insert the correct line breaks between legal paragraphs, in judicial documents of 'standard' format a new legal paragraph can be identified by a numeral or letter (in the case of lists) followed by a period. At this phase, we had to consider a number of special cases. For example, we do not insert a newline after a colon if the colon is not followed by whitespace, so as not to insert a newline in the middle of a hyperlink.

- Clean remaining lines: We remove footnotes and floating punctuation.

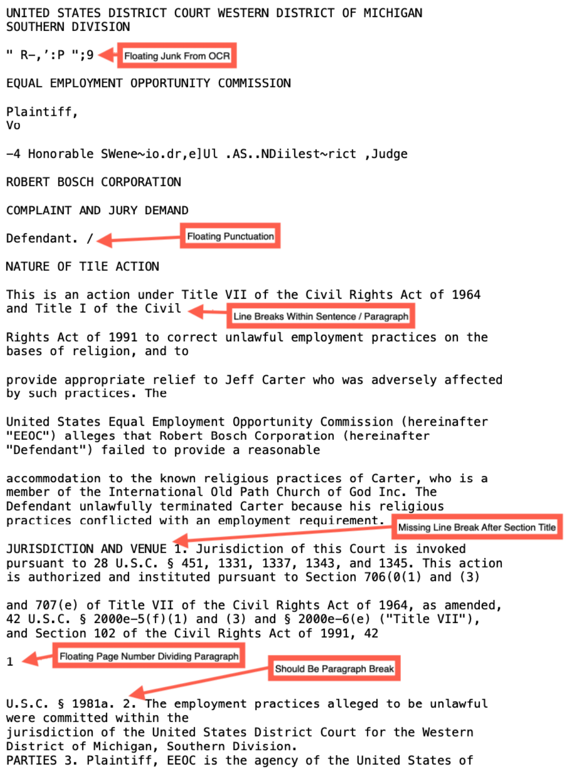
- Additional docket processing: Docket documents have a distinct format to judicial documents of other types. In particular, dockets contain tables with two columns giving the date (left), and the action taking place (right), which are not well represented in plain text format. For dockets, we remove lines consisting solely of dates (the left column of the table), and numbers at the start of lines, as this is noise from attempting to linearise the table. In the vast majority of cases, no information is lost as the corresponding date is included in the main column entry.

- Address line breaks: Removing junk information often allows us to retrieve the correct line breaks. For docket documents, this phase is different, as due to the text originally being table cells, newline characters separate sentences.

An annotated representative excerpt from a case document before and after cleaning is given in Figures 2 and 3 respectively. This displays the effectiveness of our cleaning pipeline, however we note that cleaning cannot be perfect in all cases since documents have different formats and levels of OCR noise, and we do not want to erroneously remove valid text.

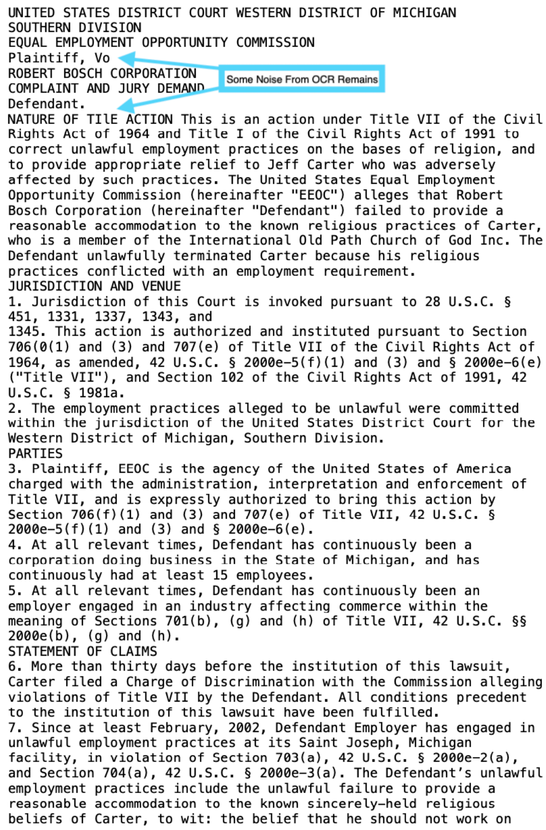
The cleaning process allows the source text to be correctly segmented into sentences and paragraphs, which is vital for subsequent stages in our methodology. The newline stages of the cleaning process allow for correct paragraph segmentation. For sentence segmentation, we use LexNLP (Bommarito et al., 2018) as this is specifically designed for legal text. Despite this, we still found that some post-processing was required to achieve the best results as certain cases were not well handled. Following (Parikh et al., 2021), we merge a sentence with the previous sentence if the previous sentence ends in an acronym (such as 'v.'), or if the current sentence begins with 'Section' (to address incorrect segmentation within legal articles). We also introduce a sentence boundary between ';' and '(' to segment long legal lists. For docket type documents, as there is no period at the end of entries in table cells, we must first divide the text into paragraphs, which correspond to each cell of the table, before applying sentence segmentation to each paragraph.

With respect to data filtering, we filter out training examples with low entity extractivity to discourage hallucination, as training examples which are unfaithful to the source text can encourage generative models to produce hallucinations (Nan et al.,



```
UNITED STATES DISTRICT COURT WESTERN DISTRICT OF MICHIGAN
SOUTHERN DIVISION

" R-,':P ";9

EQUAL EMPLOYMENT OPPORTUNITY COMMISSION

Plaintiff,
Vo

-4 Honorable SWene~io.dr,e]Ul .AS..NDiilest~rict ,Judge

ROBERT BOSCH CORPORATION

COMPLAINT AND JURY DEMAND

Defendant. /

NATURE OF TItLE ACTION

This is an action under Title VII of the Civil Rights Act of 1964
and Title I of the Civil

Rights Act of 1991 to correct unlawful employment practices on the
bases of religion, and to

provide appropriate relief to Jeff Carter who was adversely affected
by such practices. The

United States Equal Employment Opportunity Commission (hereinafter
"EEOC") alleges that Robert Bosch Corporation (hereinafter
"Defendant") failed to provide a reasonable

accommodation to the known religious practices of Carter, who is a
member of the International Old Path Church of God Inc. The
Defendant unlawfully terminated Carter because his religious
practices conflicted with an employment requirement.

JURISDICTION AND VENUE 1. Jurisdiction of this Court is invoked
pursuant to 28 U.S.C. § 451, 1331, 1337, 1343, and 1345. This action
is authorized and instituted pursuant to Section 706(0(1) and (3)

and 707(e) of Title VII of the Civil Rights Act of 1964, as amended,
42 U.S.C. § 2000e-5(f)(1) and (3) and § 2000e-6(e) ("Title VII"),
and Section 102 of the Civil Rights Act of 1991, 42

1

U.S.C. § 1981a. 2. The employment practices alleged to be unlawful
were committed within the
jurisdiction of the United States District Court for the Western
District of Michigan, Southern Division.
PARTIES 3. Plaintiff, EEOC is the agency of the United States of
```


Figure 2: Annotated representative excerpt, before cleaning process.



```
UNITED STATES DISTRICT COURT WESTERN DISTRICT OF MICHIGAN
SOUTHERN DIVISION
EQUAL EMPLOYMENT OPPORTUNITY COMMISSION
Plaintiff, Vo
ROBERT BOSCH CORPORATION
COMPLAINT AND JURY DEMAND
Defendant.
NATURE OF TItLE ACTION This is an action under Title VII of the Civil
Rights Act of 1964 and Title I of the Civil Rights Act of 1991 to
correct unlawful employment practices on the bases of religion, and
to provide appropriate relief to Jeff Carter who was adversely
affected by such practices. The United States Equal Employment
Opportunity Commission (hereinafter "EEOC") alleges that Robert
Bosch Corporation (hereinafter "Defendant") failed to provide a
reasonable accommodation to the known religious practices of Carter,
who is a member of the International Old Path Church of God Inc. The
Defendant unlawfully terminated Carter because his religious
practices conflicted with an employment requirement.
JURISDICTION AND VENUE
1. Jurisdiction of this Court is invoked pursuant to 28 U.S.C. §
451, 1331, 1337, 1343, and
1345. This action is authorized and instituted pursuant to Section
706(0(1) and (3) and 707(e) of Title VII of the Civil Rights Act of
1964, as amended, 42 U.S.C. § 2000e-5(f)(1) and (3) and § 2000e-6(e)
("Title VII"), and Section 102 of the Civil Rights Act of 1991, 42
U.S.C. § 1981a.
2. The employment practices alleged to be unlawful were committed
within the jurisdiction of the United States District Court for the
Western District of Michigan, Southern Division.
PARTIES
3. Plaintiff, EEOC is the agency of the United States of America
charged with the administration, interpretation and enforcement of
Title VII, and is expressly authorized to bring this action by
Section 706(f)(1) and (3) of Title VII, 42 U.S.C. §
2000e-5(f)(1) and (3) and § 2000e-6(e).
4. At all relevant times, Defendant has continuously been a
corporation doing business in the State of Michigan, and has
continuously had at least 15 employees.
5. At all relevant times, Defendant has continuously been an
employer engaged in an industry affecting commerce within the
meaning of Sections 701(b), (g) and (h) of Title VII, 42 U.S.C. §§
2000e(b), (g) and (h).
STATEMENT OF CLAIMS
6. More than thirty days before the institution of this lawsuit,
Carter filed a Charge of Discrimination with the Commission alleging
violations of Title VII by the Defendant. All conditions precedent
to the institution of this lawsuit have been fulfilled.
7. Since at least February, 2002, Defendant Employer has engaged in
unlawful employment practices at its Saint Joseph, Michigan
facility, in violation of Section 703(a), 42 U.S.C. § 2000e-2(a),
and Section 704(a), 42 U.S.C. § 2000e-3(a). The Defendant's unlawful
employment practices include the unlawful failure to provide a
reasonable accommodation to the known sincerely-held religious
beliefs of Carter, to wit: the belief that he should not work on
```


Figure 3: Annotated representative excerpt, after cleaning process.

2021a; Ji et al., 2023; Dong et al., 2022; Chaudhury et al., 2022; Narayan et al., 2021). While the summaries in Multi-LexSum are expertly constructed and faithful to the source documents as on the CRLC website, the OCR process means that not all documents are adequately represented by the plain text format in Multi-LexSum - for example, the dataset contains handwritten source documents for which the OCR software struggles to extract any text. Therefore, the Multi-LexSum dataset contains cases where the source text does not contain key information in the summary; these cases should be removed.

We based our filtering on verifying if the named entities in the summary occur in the source text. Firstly, in order to conduct the named entity recognition (NER), we use a state-of-the-art NER system (Barale et al., 2023) developed specifically for the legal domain in collaboration with legal professionals. The NER model was trained on human-annotated Canadian refugee law cases, fine-tuning LEGAL-BERT (Chalkidis et al., 2020). We include standard NER categories (DATE, PERSON, GPE, ORG, NORP, LAW) and the CLAIMANT_INFO legal-specific category. Additionally, we added the MONEY category from LexNLP (Bommarito et al., 2018). We manually evaluated results of the LEGAL-BERT NER systems on a subset of the validation set, studying the performance and relevance of all categories. While overall the NER system performed well, we found one common error for the GPE and ORG categories - the system included additional words between two true entities, resulting in one false entity (eg *'AT&T employee* against *AT&T Corp.'*) being returned. To solve this, we used the NLTK (Bird and Loper, 2004) part-of-speech tagger to postprocess these categories, removing words which were not nouns, adjectives, 'in', or 'of' from the entity and segmenting at the newly created boundaries.

Our filtering was based on verifying if the entities extracted from the gold standard summary appeared in the source text. However, matching named entities is nontrivial (Nan et al., 2021a), with several recurring scenarios causing difficulty:

- Dates - the same date can occur in different formats. We adopted a very optimistic approach to filter out obvious errors, however we note that this may give false positives by indicating entities are extractive when they are not. To deal with generalisations, such as ‘September 2003’ occurring in the summary while the source documents may only contain specific dates (i.e. - the day of the month is also specified), we parsed such expressions into multiple date formats and attempted to find a match in the source text for any of these formats, for any day of the month. Similarly, for expressions such as ‘early 2003’ we solely attempted to verify the year. For relative expressions such as ‘the next day’, we optimistically assumed these were valid.

- Paraphrases - for example, ‘AT&T employee’ and ‘employed by AT&T Corp.’.
- Expansion and contraction of abbreviations - for example, ‘Corporation’ and ‘Corp.’. Creating a dictionary to match all such abbreviations would be infeasible.
- Minor errors such as inconsistent spacing and punctuation.

We note that many of these issues occur due to basing our matching on an exact match of surface forms. While we considered strategies such as fuzzy string matching, we found this to lead to worse results, as for example, changing one letter is very important when referring to legal articles, but could still lead to a fuzzy string match with high confidence. Overall, while our method is not reliable at the level of individual entities, we found through manual inspection that our method suffices to filter out obviously low-quality sources. From inspection of the percentage of entities verified, summaries, court documents on the CRLC website, and source text in Multi-LexSum for a sample of cases, we removed cases where less than 75% of entities could be verified.

We found one legitimate case where summaries contained non-extractive entities: where the final sentence of the summary indicated whether the case was closed ‘as of’ the date of writing. In such cases, the date of writing was evidently not contained in the source documents. Therefore, if the last sentence of summaries in the training set contained ‘as of’, we removed this sentence so as not to encourage hallucination.

C OREO: Further Details

Formally, the OREO algorithm defines the summary-worthiness of a sentence x_i as the expectation of its associated oracle evaluation:

$$\ell'_i := \sum_{Y^*}^{Y} \mathcal{R}\left(Y^*,\ S\right)\ p\left(x_i|Y^*, D\right)\ p\left(Y^*|D, S\right) = \underset{Y^* \sim\ p(Y^*|D,S)}{E}\left[\underbrace{\mathcal{R}\left(Y^*, S\right)}_{oracle\ quality}\quad \underbrace{p\left(x_i|Y^*, D\right)}_{oracle\ membership}\ \right]$$

where $\mathcal{R}$ denotes the mean of ROUGE-1 and ROUGE-2, $D = \{x_i\}_1^m$ denotes the source text, S is the reference (abstractive) summary, and Y is the oracle summary space. The ’oracle membership’ term refers to if the oracle hypothesis Y^* is in the oracle distribution, which is a uniform distribution over the t top results of the k oracle summary hypotheses returned by beam search. The final sentence labels are given by the scaled expectation $\ell(x_i) = (\ell'_i - \bar{\ell}_{min})/(\bar{\ell}_{max} - \bar{\ell}_{min})$ (Xu and Lapata, 2022).

To obtain the OREO labels for Multi-LexSum, we set the beam size hyperparameter k to 16, and the oracle distribution hyperparameter t to 16, as in the hyperparameter search performed in Xu and Lapata (2022), these were the best parameters for the most highly compressive dataset evaluated, Multi-News. We set the summary size hyperparameter to 30 (approx. 1024 / 34) sentences, based on the mean number of tokens (34, very long tail distribution) per source sentence. However, after running OREO, in many cases fewer than 30 sentences were extracted (received a non-zero score) for a given case.

D BERT Sentence Salience Classifier: Further Details

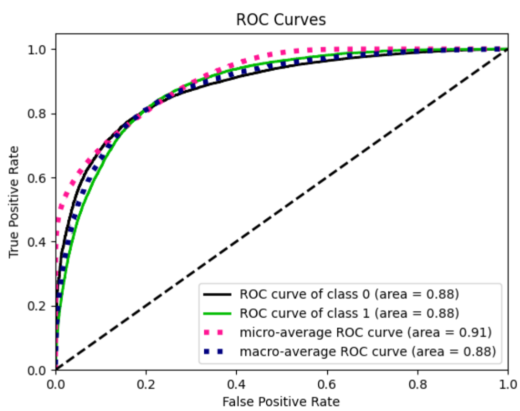


Figure 4: ROC curve for CaseLawBERT classifier.

We train the CaseLawBERT model using its Pytorch implementation in Huggingface (Wolf et al.,

2020) on a single NVIDIA RTX A6000 GPU. The model was trained for 3 epochs (Zheng et al., 2021), with a batch size of 16 and using BertAdam with a learning rate of 2e-5 and warmup of 0.01. Inputs were truncated at 128 tokens for feasibility reasons due to the huge number of sentences in the test set; we acknowledge that not truncating may lead to improved results. As output, we obtained the probability of the sentence containing salient information. As we are not working with a threshold (to construct the inputs to PEGASUS, we use a ranked list by probability) and as metrics such as accuracy, precision, and recall are not very informative for highly skewed data, we report the classifier's ROC-AUC score of 0.884 (Figure 5) - this indicates excellent (Mandrekar, 2010) performance, despite the computational considerations made.

E Details of Results: Preliminary Content Selection Experiment

Table 5 details the mean number of tokens extracted per input strategy. Figure 5 details the distributions of the ROUGE recall score per input strategy.

	OREO	BERT
Sentences	264.15	1000.78
Windows	821.31	966.10
Paragraphs	679.73	596.47

Table 5: Mean number of tokens extracted for BERT-based and OREO-based input strategies.

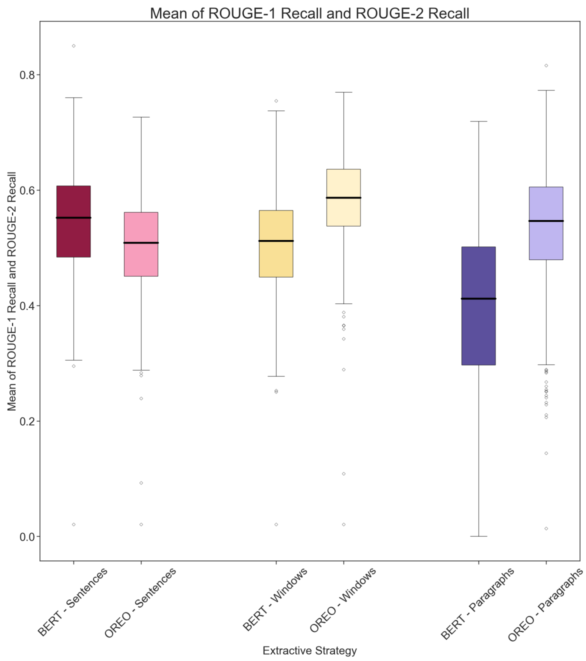


Figure 5: Distributions of ROUGE recall scores against corresponding reference summary for BERT-based and OREO-based strategies, demonstrating the difference in salient information retrieval between BERT-based and OREO-based counterparts.

F Experimental Setup - PEGASUS

All experiments were conducted on a single NVIDIA RTX A6000 GPU, using the PyTorch implementations of PEGASUS and Legal-PEGASUS available from the Huggingface (Wolf et al., 2020).

G Annotated Model Outputs

We include (Figure 6) examples of representative model outputs for two legal cases, compared with the results of our PEGASUS reproduction baseline and the gold standard summaries. Facts inconsistent with the case documents and other errors (such as assimilating information from cited cases) are highlighted in red.

	Cerda v. Restaurant Associates	City of Los Angeles v. JPMorgan Chase
Gold	On August 9, 2004, the plaintiff filed this class action lawsuit in the Eastern District of New York on behalf of all Hispanic and/or non-white refreshment concession personnel who were hired, employed or offered employment by the defendants, Restaurant Associates and RA Tennis Corp. The plaintiff alleged that the defendants engaged in unlawful employment discrimination on the basis of national origin and race respecting the assignment of positions and locations to food vendors employed at the U.S. Open Tennis Tournament. Specifically, they alleged that the defendants were deliberately and overwhelmingly assigning beer wagons to non-Hispanic white people over other demographics. The parties soon reached a settlement which was approved by U.S. District Court Judge Robert M. Levy, and Restaurant Associates agreed to pay damages. This case is closed.	On May 30, 2014, the City of Los Angeles filed a lawsuit in the U.S. District Court for the Central District of California against JP Morgan Chase Bank, under Equal Credit Opportunity Act (ECOA) and Fair Housing Act/Fair Housing Amendments Act (FHAA). The plaintiff sought damages, declaratory and injunctive relief, alleging that the defendants lending practices were predatory and intentionally discriminated against minority borrowers, and had a disparate impact on minorities. On August 6, 2015, the plaintiff field an ex parte application to stay the case pending appeal of related action. On September 2, 2015, the case was dismissed as stipulated.
Reproduction Baseline	On August 9, 2004, a restaurant worker filed a class action lawsuit in the U.S. District Court for the Eastern District of New York against the owners and managers of a restaurant in Brooklyn, New York. The plaintiff alleged that the defendants had violated the Fair Employment and Housing Act (FEH) and the Rehabilitation Act of 1973. The court granted the plaintiff's motion for preliminary injunction, and the parties agreed to a settlement. The case is closed.	On May 30, 2014, the City of Los Angeles filed a lawsuit in the U.S. District Court for the Central District of California against JPMorgan Chase & Co. and Chase Manhattan Bank. The City alleged that the banks had violated the Equal Credit Opportunity Act (ECA) and the Fair Credit Reporting Act (FCRA) by denying the City access to credit card information. The court granted the City's motion for summary judgment and granted the defendants' motion to dismiss. The parties agreed to a settlement agreement. The settlement agreement was entered into on September 2, 2015.
BERT-Windows, LegalPEGASUS	On July 20, 2009, a class action lawsuit was filed in the U.S. District Court for the Eastern District of New York on behalf of Hispanic and non-white employees of Restaurant Associates, Inc. and RA Tennis Corp. who worked at the US Open Tennis Tournament in 2003. The complaint alleged that the defendants discriminated against non-Hispanic white employees and shunted them to lower-paying positions in violation of the federal, state and City employment non-discrimination statutes. The parties reached a settlement in which the defendants established a $90,000 settlement fund to be divided pro rata among non-White employees and employees of non-U.S national origin who file claims. The settlement was approved by the Court on July 20, 2010.	On September 2, 2015, the City of Los Angeles filed this lawsuit in the U.S. District Court for the Central District of California against JPMorgan Chase Co., alleging that the bank violated the Fair Housing Act by targeting minority borrowers for predatory loan terms and increasing the costliness of non-conventional loans at the expense of minority borrowers. The case is one of four related cases brought by the City against large lending institutions, including Wells Fargo, Citigroup Inc., and the California Department of Housing and Urban Development. In 2012, the City entered into a global settlement agreement with the government pursuant to which JPMorgan agreed to pay $753 million into a settlement fund for distribution 10 to qualified borrowers; and (b) provide an additional $1.2 billion to foreclosure prevention actions. The City's damages include lost tax revenues and the need to provide 21 increased municipal services.

Restaurant Associates and RA Tennis Corp are food vendors at the U.S. open tennis tournament, not a physical restaurant, and owners and managers were not directly involved - this demonstrates a faithfulness problem not related to entities.

Throughout, dates included occur in the source text, but appear in the incorrect context in the generated summary.

Information following ‘In 2012’ is related to a cited case, not the current case, and contains artefacts

Figure 6: Annotated examples of representative model outputs for two cases, with facts inconsistent with the case documents and other errors (such as assimilating information from cited cases) highlighted in red.)