# Reconceptualising Automatic Text Summarisation Evaluation: Evidence from Long Legal Texts

**Charlie Brown**
UCL
Gower Street
London
WC1E 6BT

**Rory Creedon**
UCL
Gower Street
London
WC1E 6BT

**Siddartha Nath**
UCL
Gower Street
London
WC1E 6BT

**Tik Nang Tsoi**
UCL
Gower Street
London
WC1E 6BT

## Abstract

Legal documents, such as contracts, case law and legislation, are often very long by nature. Reading, understanding and conducting analysis of these long documents can be a time-consuming and challenging task. This makes legal documents an area ripe for the use of automatic text summarisation (ATS). Within Natural Language Processing (NLP) literature, the use of ATS has largely been evaluated using metrics such as ROUGE (Lin, 2004), which compare summarised text to a reference summary produced by a human. However, there are two key problems with using reference-based summary metrics such as ROUGE - firstly, they are limited by the quality of the human summary and secondly, they do not capture how useful the summarised text is for a range of different human tasks. To combat this, we propose an alternative framework for evaluating the effectiveness of summarised text for inference. Our results [1] indicate that the widely used neural network systems, specifically transformers, (Barbella et al., 2021), fail to outperform older graph-based summary techniques on inference tasks despite higher ROUGE scores (Koh et al., 2022).

## 1 Introduction

Natural Language Processing (NLP) has become increasingly prevalent in the field of legal research and analysis (Zhong et al., 2020). Legal professionals must be able to read and analyse a vast amount of text, which can often be time-consuming and tedious. It can provide an efficient solution for extracting relevant information and key insights from legal texts, making it an invaluable tool for legal practitioners in processing legal documents.

One of the key applications of NLP in legal text processing is automatic text summarisation (ATS) (Kanapala et al., 2019). ATS enables the creation of condensed versions of texts that include only the most important information, allowing legal professionals to quickly identify the key details in a document. The quality of summarisation is commonly evaluated using intrinsic evaluation metrics such as the ROUGE score, which measure the similarity between the generated summary and a reference summary (written by a human), as opposed to extrinsic evaluation metrics, which evaluate the quality of summaries for a given task. There are indications that intrinsic metrics may be overused - in a previous study, it was reported that 100% of papers when introducing new summarisation models at the Computational Linguistics Conferences in 2021 use ROUGE, with 69% using it exclusively (Gehrmann, 2022).

However, reference-based summary metrics are not without their limitations. Firstly, reference summaries are highly subjective, as they are written by humans and may omit essential details and nuances that could be critical in legal texts. Secondly, producing summaries that are similar to human-generated reference summaries does not necessarily mean that the summaries will be useful for given downstream tasks. Belz and Gatt (2008) found that there was no significant correlation between the intrinsic evaluation of summaries (i.e. using reference summaries) and the extrinsic evaluation of summaries (i.e. inference tasks). It is therefore crucial to take into account both intrinsic and extrinsic metrics when evaluating the quality of a summary.

There have been some attempts to incorporate more extrinsic evaluation measures such as question answering (QA) (Eyal et al., 2019).

[1] Our implementation and results can be accessed via https://github.com/rorycreedon/comp0087_assignment

However, these do not scale well to longer texts with longer summaries (see section 2.2). As the dataset used in this paper features long texts, we propose an alternative extrinsic evaluation measure - the performance of a separate machine learning model on a range of inference tasks, using the summarised text as an input.

To conduct a comprehensive study, we chose the ECHR dataset (Chalkidis et al., 2019), which encompasses three distinct inference tasks, including a binary classification of whether a European Court of Human Rights (ECHR) article was breached, a multi-class classification of which article was breached and a regression on the importance score of the case assigned by the ECHR.

For inference, we selected LEGAL-BERT (Chalkidis et al., 2020), a variant of the BERT transformer (Devlin et al., 2018) for the legal domain. LEGAL-BERT has been specifically pre-trained on legal documents from multiple jurisdictions and has demonstrated superior performance on several legal tasks (Chalkidis et al., 2020).

## 2 Related Work

### 2.1 Automatic Text Summarisation

Research in ATS began in the 1950s, with the first approach conducted via statistical analysis, focusing on word frequency and distribution a relative measure of text significance (Luhn, 1958). With improvements in CPU/GPU architecture and understanding of NLP via machine learning, the development of new summarisation models commenced.

In the early 2000s, the introduction of graph-based models, such as TextRank and LexRank, greatly improved automatic summarisation by using graph theory to identify important sentences in a document (Mihalcea and Tarau, 2004; Erkan and Radev, 2004). These models were based on the idea of treating a document as a graph, with sentences as nodes and edges representing relationships between them. The issues with these methods are scalability and the lack of linguistic semantics.

Heading into the next decade, the development of deep learning models such as Recurrent Neural Networks (Shini and Kumar, 2021) and Transformers (Beltagy et al., 2020) led to significant improvements in summarisation performance (Koh et al., 2022). The transformer architecture is composed of multiple layers with two sub-layers: a self-attention sub-layer and a feed-forward sub-layer. The self-attention sub-layer computes attention weights for input tokens based on their relation to all other tokens, while the feed-forward sub-layer applies non-linear transformations to the outputs of the self-attention sub-layer. As a result of this architecture, these models are able to capture more complex relationships between sentences and thus provide a better basis for inference.

There are two main approaches in ATS: extractive and abstractive summarisation. Extractive summarisation (Wong et al., 2008) selects phrases or sentences directly from the source text to create a summary, whilst abstractive summarisation (Khan and Salim, 2014) conveys the same ideas in the source document using different words and phrasing. There exist both abstractive and extractive summarisation methods using both more traditional graph-based methods and neural networks, however, abstractive methods using graph-based methods tend to have very low linguistic quality (Khan and Salim, 2014). Therefore, abstractive graph-based methods are out of scope in this paper.

### 2.2 Evaluation Measures

Evaluating the performance of summaries is a challenging task that can be divided into two approaches - intrinsic metrics and extrinsic metrics. In intrinsic evaluation, the quality of the summary is judged based on an analysis of the summary. Typically, this takes the form of a comparison to a 'reference' summary - a summary generated by a human (ideally with strong knowledge of the relevant domain). In extrinsic evaluation, the quality of the summary is judged based on helpful the summary is for a given task. This can take the form of question-answering (QA) or, in the case of this report, conducting inference based on the summary.

As mentioned in section 1, ROUGE (Lin, 2004) is the de-facto 'gold standard' of summarisation evaluation. It is an intrinsic metric, which compares the lexical overlap between the generated summary and a reference summary. There are five different variants of ROUGE. In equation 1, we

define ROUGE-N, which measures the overlap of n-grams (different ROUGE metrics measure different overlaps), where $r_i \in R$ is the reference summary $i$ in the set of reference summaries $R$, $g_n$ is an n-gram, $C(g_n)$ is the number of n-grams in the generated summary and $C_m(g_n)$ is the number of n-grams in both the generated summary and the reference summary.

$$\text{ROUGE-N} = \frac{\sum_{r_i \in R} \sum_{g_n \in r_i} C_m(g_n)}{\sum_{r_i \in R} \sum_{g_n \in r_i} C(g_n)} \quad (1)$$

ROUGE is not, however, without limitations. It suffers from the inclusion of false matches between tokens that do not convey the same information, thereby reducing its accuracy. To overcome this limitation, recent methods such as MoverScore (Zhao et al., 2019) and BERTScore (Zhang et al., 2020) have been developed, which evaluate the similarity of tokens based on their contextual embeddings, thereby improving the accuracy of evaluation.

Extrinsic metrics, by nature, are more difficult to generalise as they measure the quality of summarised text for a given task. There are a variety of different techniques for extrinsic evaluation, but most revolve around question-answering (QA). This can involve reading comprehension, where a human answers multiple choice questions based on both the original text and the summary and relevance assessments, where topics summarised are classified based on their relevance (Inderjeet, 2009).

Some recent papers (Eyal et al., 2019; Deutsch et al., 2021) use QA as an extrinsic evaluation metric, however, the use of QA does not scale well to long texts. This is because the number and the complexity of questions (which are typically written by a human) needed to evaluate summaries will increase exponentially with the length of the text being summarised. Whilst there is some literature on datasets with QA tasks with long inputs (Pang et al., 2022), it is typically a very long and expensive procedure to build such datasets. In this paper, we present an alternative to large QA datasets - the use of a range of inference tasks as an extrinsic evaluation method. Typically, data for inference tasks (such as predicting if there has been a violation of an ECHR article) exists in the real world, whereas QA datasets tend to be purposefully crafted. As such, our framework can be used where QA data is not available.

## 3 Methodology

The framework that we propose for evaluating the performance of summarised text on various inference tasks is shown in Figure 1. The steps taken in the methodology are outlined below:

1. **Data Collection**: Data from the ECHR dataset was downloaded and pre-processed. This involved label encoding each of the violated and allegedly violated articles.

2. **Data Preparation**: Each of the texts in the ECHR was summarised, where a different summary was generated for each text using each of the summarisation models (i.e. TextRank, GL-LSTM, etc.). For certain models, some texts had to be truncated before summarisation (i.e. for the LED Longformer, texts were truncated to 16,843 tokens before summarisation).

3. **Modelling**: Three different inference tasks were performed on the summarised texts, all using LEGAL-BERT: (1) binary classification (determining whether any human rights article was violated), (2) multi-label classification (identifying which human rights articles were violated, if any), and (3) regression (assessing the importance of a case on a scale of 1 to 4). Hyperparameter optimisation was performed on the learning rate and batch size. Using these hyperparameters, a final model was trained for each of the inference tasks.

4. **Model Evaluation:** The performance on each inference task using each of the summarised texts was evaluated using the held out testing set. To address the class imbalance in the classification tasks (see section 5.1), we selected weighted F-1 scores as the evaluation metrics. For the regression task, we opted to use mean absolute error, as recommended by Chalkidis et al. (2020).

## 4 Models

### 4.1 Long Text Summarisation

Summarisation models for this paper were selected based on a recent survey paper on long document summarisation (Koh et al., 2022), which are broken

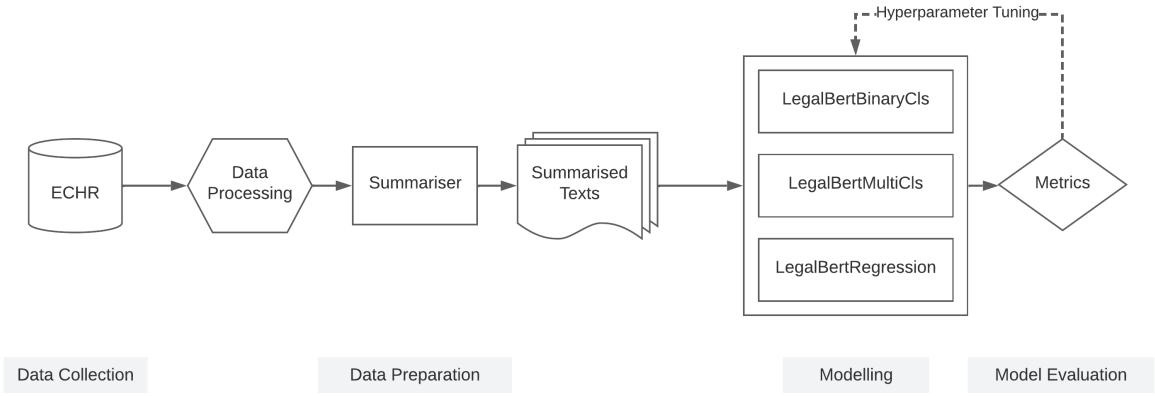


**Figure 1:** End-to-end pipeline of methodology.

down into three categories - graph-based methods, neural network-based extractive methods and neural network-based abstractive methods. Given the length of the texts in our dataset (see Figure 2), only models that could handle long input sequences were included.

#### 4.1.1 Graph-Based Models

**TextRank:** An unsupervised graph-based extractive summarisation algorithm that employs a variant of PageRank (Mihalcea, 2004) to find the most important lexical units in a document, based on relevant keywords (Mihalcea and Tarau, 2004). The algorithm models a document as a graph, where sentences serve as nodes, and weights the edges between them using a similarity function that considers the shared content between them,

$$Sim\left(S_i, S_j\right) = \frac{|\{w_k \mid w_k \in S_i \& w_k \in S_j\}|}{\log\left(|S_i|\right) + \log\left(|S_j|\right)} \quad (2)$$

given two sentences, $S_i, S_j$, represented by a set of $n$ words that in $S_i$ are represented as $S_i = w_1^i, w_2^i, \ldots, w_n^i$. The importance of each node is measured based on its centrality within the graph, and PageRank is applied over the adjacency matrix to rank the nodes and select the most significant sentences.

**LexRank:** A variation of TextRank that incorporates a Term Frequency-Inverse Document Frequency (TF-IDF) model instead of the baseline word frequency term used in the bag of words model (Erkan and Radev, 2004). In addition, LexRank uses a TF-IDF modified cosine similarity function to determine sentence similarity when building the adjacency matrix,

$$\cos(x, y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} \left(idf_w\right)^2}{\sqrt{\sum_{x_i \in x} \left(tf_{x_i x} idf_{x_i}\right)^2} \sqrt{\sum_{y_i \in y} \left(tf_{y_i y} idf_{y_i}\right)^2}} \quad (3)$$

**Graph-Reduction:** A graph-based approach similar to Text/LexRank, however, to reduce the complexity of the graph while preserving its meaning, predicate-argument mapping and normalisation are used to reduce the adjacency matrix (Fabish, 2014). The sentence salience in this method is determined by summing the weights of its edges to other sentences after the graph has been reduced.

#### 4.1.2 Extractive neural-network models

**GL-LSTM:** A model that considers both the global context of the entire document and the local context within the current topic (Xiao and Carenini, 2019). Inspired by the natural topic-oriented structure of human-written long documents, the model incorporates discourse information, such as section structure, by encoding the section-level and document-level representation (using LSTM-Minus) into each sentence, significantly improving model performance. To create the extractive summary, each sentence is visited. The model consists of three main components: the sentence encoder, the document encoder, and the sentence classifier.

**BERT-Extractive-Chunking:** In order to tackle the 512 token input limit of pre-trained transformer models such as BERT, chunking approaches have been proposed (Joshi et al., 2019). This involves dividing the input document into overlapping or non-overlapping chunks of no more than 512 tokens,

processing each chunk separately and then combining the resulting summaries. Bert-Extractive-Chunking, utilises BERT to generate text embeddings for each chunk of the document independently and then applies K-Means clustering to identify sentences closest to the centroid for summary selection (Miller, 2019).

#### 4.1.3 Abstractive neural-network models

**Long-T5:** An extension to the original large-scale encoder-decoder transformer-based language model T5 (Raffel et al., 2020) developed by Google, for its ability to handle long input sequences of up to 16,384 tokens. By incorporating transient-global attention and local attention mechanisms and utilising attention sparsity patterns, the model can efficiently handle long input sequences (Guo et al., 2021). This model has been pre-trained in a text-to-text denoising generative setting on a large corpus of the English language and subsequently fine-tuned on summarisation datasets to improve its performance in the domain of text summarisation.

**LED-Longformer:** A modified version of the Longformer model that adopts an encoder-decoder architecture designed for sequence-to-sequence learning with an attention mechanism that scales linearly with the input (Beltagy et al., 2020). Unlike Longformer, which only has an encoder architecture, LED Longformer uses an efficient local and global attention pattern on the encoder network, enabling it to handle long document sequence-to-sequence tasks such as summarisation for up to 16,384 tokens. The decoder component of LED Longformer employs a full self-attention mechanism that attends to both the encoded tokens and previously decoded locations, allowing it to generate a summary that captures the salient information of the input document.

### 4.2 Inference tasks

For the inference tasks, the same pre-trained model (LEGAL-BERT) was used for all three tasks, which is detailed below.

**LEGAL-BERT-SMALL:** A light-weight version of LEGAL-BERT which trains four times faster while still maintaining comparable performance (Chalkidis et al., 2020). LEGAL-BERT-SMALL (referred to LEGAL-BERT in this paper) was fine-tuned for each inference task. The model architecture was adjusted with a fixed dropout rate of 0.2, followed by a linear layer with the number of required outputs. For the classification tasks, a sigmoid layer was added at the end.

## 5 Experiments

### 5.1 Dataset

The ECHR dataset consists of roughly 11.5k court cases sourced from the ECHR public database (Chalkidis et al., 2019), with each case containing the relevant facts leading to the judgement.

Experiments were conducted on the training, validation and test splits that were created for the dataset. Note that the dataset offers both a non-anonymised and anonymised version to encourage further studies on fairness and biases. As this is not the major focus of this paper, the non-anonymised version was chosen.

The test, train and validation set all exhibit a highly right-skewed distribution in terms of text length, with a minimum length of 61 and a maximum length of just over 30k (see Figure 2).

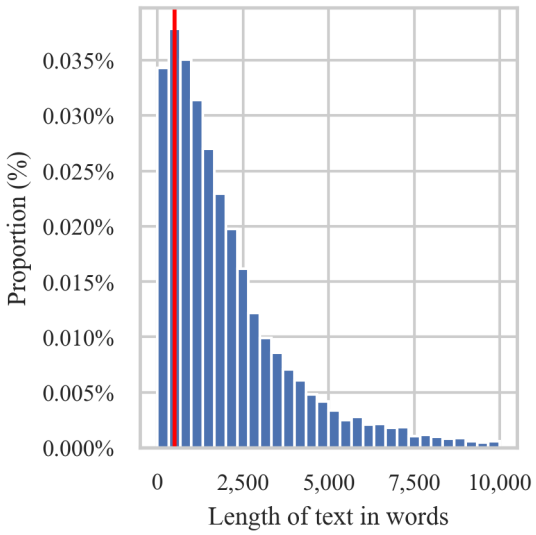


**Figure 2:** Histogram of length of texts in the ECHR dataset.

Notes: (i) Excludes texts longer than 10,000 words. (ii) The red line represents the maximum input length of LEGAL-BERT (512 tokens).

We also note that the test set contains a higher proportion of cases that contain violations (66%) than in the training set (50%). Additionally, we find that there are altogether 23 labels for the multi-class classification, but only 19 of these labels are present in both the training and test sets. Note also

11 labels appear fewer than 50 times in the training set. As mentioned in section 3, weighted F-1 scores are used as a metric for the classification tasks to account for this issue.

### 5.2 Text Summarisation

The model used for inference, LEGAL-BERT, takes in a maximum input length of 512 tokens. As discussed in section 4.1, this is typical of many modern neural-network-based models. The generated summaries from each summarisation model have a maximum length of 512 words. For the graph-based models, the number of sentences, as opposed to the number of words, is specified when running the summarisation model. Ten sentence-long summaries were generated. If this exceeded 512 words, the summarisation model was run with one fewer sentence until the final summary had fewer than 512 words.

### 5.3 Baseline Testing

One naive approach for dealing with long documents using transformers is to truncate the original texts down to the maximum size the model can handle (Limsopatham, 2021; Mamakas, 2022). Truncating the original text to LEGAL-BERT's maximum input length, 512 tokens, was used as our baseline summarisation technique, which is referred to in our results as "Truncation". Additionally, to control for the possibility that much of the relevant information is contained at the start of the text, we additionally include a second baseline model, where only the final 512 tokens are used. This is referred to in our results as "Truncation (End)".

### 5.4 Hyperparameter Settings

Devlin et al. (2018); Chalkidis et al. (2020) suggest clear hyperparameter ranges to fine-tune BERT and LEGAL-BERT. We adapted their strategy based on our available computing resources, and conducted a grid search on learning rate $\in \{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 4 \times 10^{-5}\}$ and batch size $\in \{4, 8, 16\}$. We noted that the three tasks tend to converge at different rates, i.e. based on training and validation losses, the binary classification models tended to overfit after 1 epoch while the multi-class classification models often continued to underfit even after 4 epochs. As a result, we used early stopping on the validation loss with patience of 1 and a maximum number of 5 training epochs. While the validation loss is a good indicator of convergence, it is possible that a lower validation loss does not necessarily indicate better predictive performance. To this end, we saved the optimised models based on their validation scores, which correspond to the evaluation metrics, F-1 score and MAE, as discussed in section 5.1.

## 6 Results

The results from the regression task can be seen in Figure 3 and from the classification tasks in Figure 4. In both charts, the blue bars represent abstractive neural network models, the red bars represent extractive neural network models, the green bars represent graph-based models and the purple bar represents the baseline - no text summarisation.

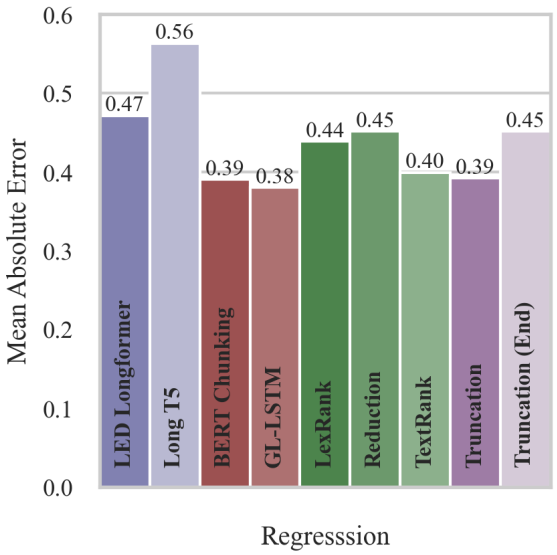


**Figure 3:** Mean Absolute Error for regression task.

The results in Figures 3 and 4 appear to support the findings of Belz and Gatt (2008), that there is no significant correlation between performance on intrinsic and extrinsic evaluation metrics. There is no family of models (or even a single model) that consistently outperforms the others, despite how different families of models tend to perform using ROUGE scoring (Koh et al., 2022).

Moreover, the results show no strong correlation between performance between the three inference tasks. The correlation matrix for the summarised texts (excluding truncation) is shown in Table 1. The lack of a strong correlation between

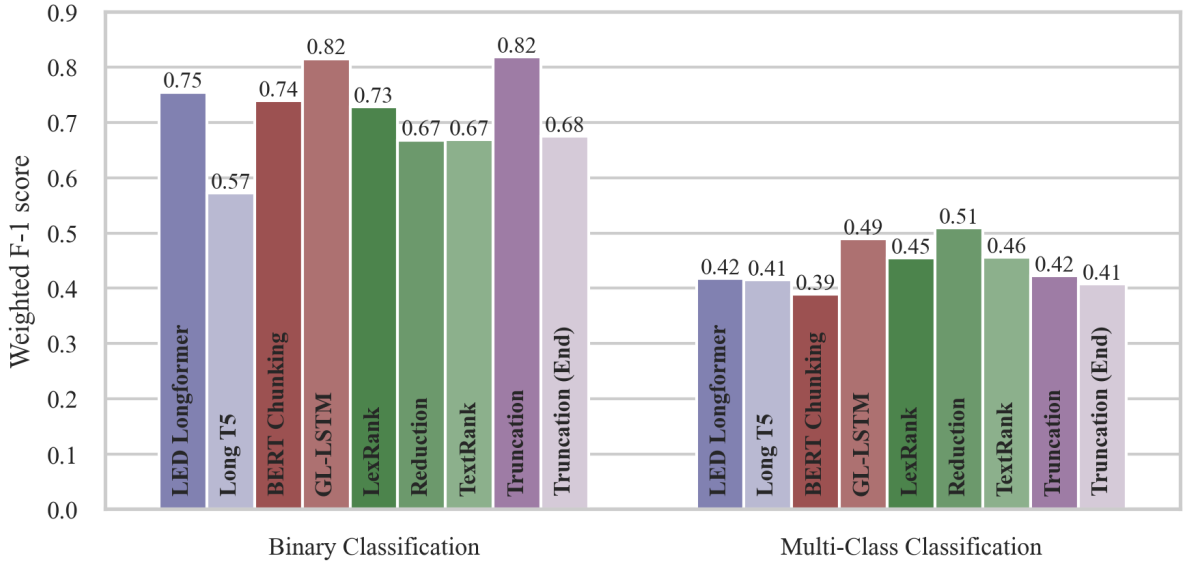


**Figure 4:** Weighted F-1 scores for classification tasks

| | Binary | Multi | Regression |
|---|---|---|---|
| **Binary** | 1.00 | 0.15 | -0.74 |
| **Multi** | | 1.00 | -0.26 |
| **Regression** | | | 1.00 |

**Table 1:** Correlation matrix of performance on each inference task.

Note: Excludes truncation summaries.

performance on the different inference tasks shows that the tasks are very distinct, meaning that a set of summaries could perform well on one task but not on another (i.e. LED Longformer performs well on binary classification, but not on regression). This shows that the summaries do not perform similarly on different tasks, meaning that, on the whole, they are unlikely to perform well on a variety of different tasks.

Finally, we note that the performance of truncation achieves the best performance for the binary classification task along with comparable performance in the multi-class classification and regression tasks. Despite the fact that the vast majority of texts in the ECHR dataset are much longer than 512 words (see Figure 2), it exhibited superior performance to modern, complex models.

## 7 Discussion

The results of this paper show that there does not appear to be a significant relationship between intrinsic and extrinsic evaluation measures for the automatic summarisation of long legal texts. On the surface, this appears to be a somewhat surprising result. Despite the recent advancements in NLP models, modern neural network-based techniques fail to outperform traditional graph-based models on the extrinsic evaluation tasks in our experiments.

One possible hypothesis for the failure of neural network models to consistently outperform more traditional graph-based models is that neural network summarisation models are typically trained using reference summaries (at the token level). Therefore, neural network-generated summaries are optimised on their similarity with the reference summaries. This could explain why neural network-generated summaries often score highly on intrinsic measures such as ROUGE and do not replicate this high performance in our inference tasks, which are extrinsic measures. Graph-based models, however, are not optimised on either intrinsic or extrinsic metrics, potentially explaining why graph-based models and neural network models have similar extrinsic metrics but neural network models outperform when it comes to intrinsic metrics such as ROUGE.

The comparatively strong performance of trun-

cation is likely due to a large portion of the text relevant to the inference tasks being at the start of the texts. When using only the last 512 words ("Truncation (End)" in Figures 3 and 4), performance decreases significantly in the binary classification and regression tasks.

## 8 Conclusion and Future Work

The key result of this paper is that, for long (legal) text summarisation, higher scores on intrinsic summarisation metrics, such as ROUGE, do not necessarily lead to higher scores on extrinsic summarisation measures. Given the dominance of intrinsic, reference-based evaluation metrics (Gehrmann, 2022), this finding leads to the conclusion that extrinsic metrics, such as QA and inference tasks, should be considered when evaluating the quality of text summarisation. This is particularly the case for long texts, where the reference summary is more subjective than in shorter texts, as different humans may have different ideas of the ideal summary and the summary may be used for different purposes.

There is also potential for further work in the empirical analysis of intrinsic and extrinsic summarisation methodologies. In our dataset, we did not have reference summaries available, so we could not directly compare ROUGE scores to performance on our inference tasks. A direct comparison would add significant weight to our findings. The use of inference tasks akin to those used in this paper is another area for further research, as the performance on a range of extrinsic measures gives an idea of the quality of generalisable information contained within the summary, regardless of how 'human-like' generated summaries are. Additionally, using multiple inference models with different architectures (as opposed to just using LEGAL-BERT), multiple datasets (additional legal datasets could include the Caselaw Access Project (Caselaw Access Project, 2008), Pile of Law (Henderson et al., 2022), and the US Supreme Court Database (Spaeth et al., 2022)) and datasets from different domains would increase the robustness of our results.

## References


Marcello Barbella, Michele Risi, and Genoveffa Tortora. 2021. A Comparison of Methods for the Evaluation of Text Summarization Techniques. *10th International Conference on Data Science, Technology and Applications*.

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

Anja Belz and Albert Gatt. 2008. Intrinsic vs. Extrinsic Evaluation Measures for Referring Expression Generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200, Columbus, Ohio.

Caselaw Access Project. 2008. Caselaw Access Project. `https://case.law/`. Last accessed 2nd April 2023.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. *arXiv:1906.02059*.

Ilias Chalkidis, Prodromos Malakasiotis, Manos Fergadiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. *arXiv:2010.02559*. Version 1.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. *Transactions of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional transformers for Language Understanding. *arXiv:1810.04805*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question Answering as an Automatic Evaluation Metric for News Article Summarization. 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 3938–3948.

Adam Fabish. 2014. Reduction. Last accessed 3rd April 2023.

Sebastian Gehrmann. 2022. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. LongT5: Efficient Text-To-Text Transformer for Long Sequences. *arXiv:2112.07916*. Version 2.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset.

Mani Inderjeet. 2009. Summarization evaluation: an overview. In *Proceedings of the NTCIR Workshop*, volume 2.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for Coreference Resolution: Baselines and Analysis. 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)*. Association for Computational Linguistics.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text Summarization from Legal Documents: A Survey. *Artificial Intelligence Review*, 51(3):371–402.

Atif Khan and Naomie Salim. 2014. A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology*, 59(1):64–72.

Huan Yee Koh, Jiaxin Ju, Ming Li, and Shirui Pan. 2022. An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics. *ACM Computing Surveys*, 55(8):1–35.

Nut Limsopatham. 2021. Effectively Leveraging BERT for Legal Document Classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hans Peter Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*.

Dimitris Mamakas. 2022. Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL interactive poster and demonstration sessions*, pages 170–173.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Derek Miller. 2019. Leveraging BERT for Extractive Text Summarization on Lectures. *arXiv:1906.04165*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question Answering with Long Input Texts, Yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *axXiv:1910.10683*.

R. Subha Shini and V.D. Ambeth Kumar. 2021. Recurrent Neural Network based Text Summarization Techniques by Word Sequence Generation. *6th International Conference on Inventive Computation Technologies (ICICT)*.

Harold J. Spaeth, Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2022. 2022 Supreme Court Database, Version 2022 Release 01.

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 985–992.

Wen Xiao and Giuseppe Carenini. 2019. Extractive Summarization of Long Documents by Combining Global and Local Context. *arXiv:1909.08089*.

Tianyi Zhang, Felix Wu Varsha Kishore, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. *arXiv:1909.02622*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Yi-Li Lin, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. *arXiv:2004.12158*.