PreSumm: Predicting Summarization Performance Without Summarizing

Steven Koniaev*, Ori Ernst*, and Jackie Chi Kit Cheung

Mila – Quebec Artificial Intelligence Institute McGill University
{oriern}@gmail.com
{steven.koniaev@mail., jackie.cheung@}mcgill.ca

Abstract

Despite recent advancements in automatic summarization, state-of-the-art models do not summarize all documents equally well, raising the question: why? While prior research has extensively analyzed summarization models, little attention has been given to the role of document characteristics in influencing summarization performance. In this work, we explore two key research questions. First, do documents exhibit consistent summarization quality across multiple systems? If so, can we predict a document's summarization performance without generating a summary? We answer both questions affirmatively and introduce PreSumm, a novel task in which a system predicts summarization performance based solely on the source document. Our analysis sheds light on common properties of documents with low PreSumm scores, revealing that they often suffer from coherence issues, complex content, or a lack of a clear main theme. In addition, we demonstrate PreSumm's practical utility in two key applications: improving hybrid summarization workflows by identifying documents that require manual summarization and enhancing dataset quality by filtering outliers and noisy documents. Overall, our findings highlight the critical role of document properties in summarization performance and offer insights into the limitations of current systems that could serve as the basis for future improvements.

1 Introduction

Recent years have witnessed a remarkable proliferation of summarization models, with many achieving impressive performance on widely used benchmarks. These models, often rooted in large-scale language modeling, represent a significant leap forward in natural language processing capabilities.

Despite these advancements, not all documents are equally well summarized by these models.

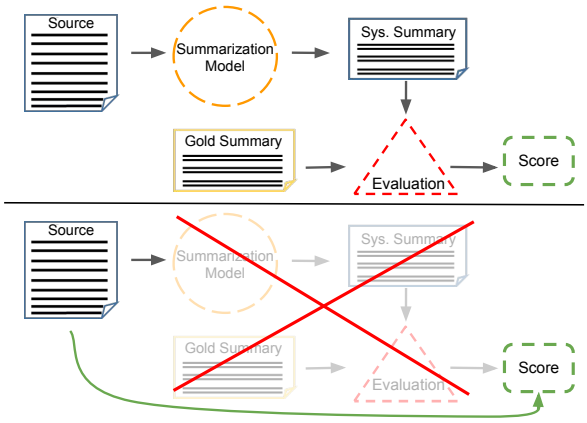

Figure 1: An illustration comparing the traditional evaluation process (top) with our approach (bottom).

Some documents yield better summaries, while others do not. As many summarization models share inherent design principles, operational mechanisms, and therefore limitations, we hypothesize that documents that are 'difficult' for one system to summarize tend to also be 'difficult' for other systems. If this hypothesis holds, it suggests that certain intrinsic properties of source documents systematically hinder summarization performance. While prior work has extensively analyzed the characteristics of summarization *models* (Goyal et al., 2022; Zhang et al., 2024), there is a lack of research on how *document* properties influence performance across systems.

To bridge this gap, we introduce the PreSumm task. In this task a system should predict a document's performance in future summarization tasks based only on the document, without generating a summary. An illustration of our approach is presented in Figure 1. Success in this task suggests that we can distinguish in advance documents that most models summarize successfully, from those where performance falters.

Analyzing such a successful PreSumm model can reveal critical source-based features that chal-

* Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2025, pages 18289–18305

lenge current systems. Gaining deeper insights into these document-level limitations could not only allow targeted improvements in summarization models but also enable strategic pre-processing document edits to facilitate the summarization process.

Additionally, PreSumm offers several practical benefits that could improve the efficiency and effectiveness of summarization workflows. One key advantage is its potential for hybrid systems where humans can focus their attention on more difficult summary cases. For example, organizations often evaluate their output by relying on automatic metrics or manual human evaluation, which can reveal poor-quality summaries that require further manual summarization—an expensive and time-consuming process. By leveraging PreSumm to identify low-performing documents in advance, organizations can prioritize these cases for manual summarization or decide to opt-out before engaging in costly summarization workflows, thereby saving valuable resources and optimizing operational efficiency.

Furthermore, a PreSumm model may identify outliers and noisy documents in the dataset whose removal could enhance outcomes. An appealing example is multi-document summarization task, where one or more problematic documents might degrade the overall quality of the output. By filtering out such documents, PreSumm can help ensure more consistent, high-quality results across summarization tasks. These practical applications highlight PreSumm's potential as a valuable tool for improving both the cost-effectiveness and performance of summarization systems.

To enable this approach, we first confirm our initial hypothesis by demonstrating a moderate correlation between document performance ranking across various systems (Section 4.2). Next, we train several models for the PreSumm task, relying solely on the source document (Section 5). Surprisingly, the supervised PreSumm models exhibit a stronger correlation with human evaluations not only compared to other baselines but also surpassing conventional automatic metrics, despite these metrics having access to the system summary. These findings also support our hypothesis that certain global document features influence summarization performance across different systems.

We also show promising results in two downstream tasks as an extrinsic evaluation. First, leveraging PreSumm predictions outperforms baselines in detecting documents that require manual summarization (Sec. 6.1). Second, we improve summarization of multi document sets by filtering out documents with low PreSumm scores (Sec. 6.2).

A deeper analysis reveals that our best PreSumm model assigns low scores to documents with coherence problems, complex content, and without a clear main theme (Section 7). Overall, to the best of our knowledge, we are the first to focus on document properties that hinder modern summarization systems. By providing deeper insights into these limitations, PreSumm establishes a strong foundation for targeted advancements and future improvements in summarization research.[1]

2 Related Work

Our work draws significant inspiration from PreQuEL (Don-Yehiya et al., 2022), which introduces a similar approach centered on machine translation. PreQuEL aims to predict translation system performance based solely on the source text. While our motivation and general methodology align with theirs, to the best of our knowledge, we are the first to apply this approach to text summarization. By leveraging PreSumm, we have enhanced summarization-specific applications and explored features more relevant to summarization, making our contributions distinct and novel.

While recent approaches (e.g., (Vig et al., 2022; Zhang et al., 2024) focused on embedding representation and did not leverage explicit source-based features, in the past, many summarization systems relied on explicit document-based features. These systems focused primarily on selecting key source sentences for inclusion in the summary—a task often framed as a classification problem. These features ranged from the presence of cue phrases (Gupta et al., 2011; Kulkarni and Prasad, 2010), the inclusion of numerical data (Prasad et al., 2012; Abuobieda et al., 2012), sentence length (Fattah and Ren, 2009; Abuobieda et al., 2012), and sentence position (Barrera and Verma, 2012; Fattah and Ren, 2009; Abuobieda et al., 2012; Li et al., 2016), to discourse structure (Louis et al., 2010), among others. While these studies employed such features to generate summaries, one may suggest a potential link between these features and the performance of summarization models. However, this property was not explicitly examined. In contrast, our work explores *system-independent document-level* features that impact the performance of mod-

[1] Our best model and its predictions over test data is publicly available at https://github.com/oriern/PreSumm.

ern summarization models in both *abstractive* and *extractive* modes.

3 Task Definition

Our task is to predict the average performance of summarization systems on a given document, using only the document as input. Averaging performance across multiple systems can reveal key properties of the document itself, while minimizing the influence of system-specific variability or noise. Formally, let $\mathcal{D}$ represent a corpus of N text passages, where each document is denoted as d_i. PreSumm aims to estimate the average quality of summaries generated by multiple systems for each document.

Specifically, consider M systems, denoted as $s_1, \ldots, s_M$, where system s_i produces a summary for document d_j that is scored as $S_{i,j}$. The goal of PreSumm is to first be able to generate a function $f : \mathcal{D} \rightarrow \mathbb{R}$ that predicts the average score assigned to the summaries of document d_j across all systems:

$$d_j^* = \frac{1}{M} \sum_{i}^{M} S_{i,j} \quad (1)$$

This score, d_j^*, reflects the average quality of the summaries generated by different systems for document d_j. We could leverage this score to rank the documents by their potential performance.

To evaluate the model's performance on ranking documents, we measure the correlation between the predicted scores and the gold-standard scores, following standard practices for assessing summarization metrics (Fabbri et al., 2021).

4 Dataset and Preliminary Analysis

4.1 The RoSE Dataset

The RoSE dataset (Liu et al., 2023) introduces a new method for annotating summarization datasets, which improves annotator agreement via *Atomic Content Units* (ACUs). The protocol tasks an annotator to convert a reference summary into atomic factual statements, then to compare the generated summary to these ACUs. The ACU score is defined as $f(s, A) = \frac{|A_s|}{|A|}$ where $|A|$ is the total number of ACUs for a given reference summary and $|A_s|$ is the number of matched ACUs of the generated summary with respect to the reference.

The authors manually evaluated over 22,000 summary-level annotations across 2,500 documents summarized by 28 top-performing systems on three datasets (CNN/DailyMail (Nallapati et al., 2016), XSum (Narayan et al., 2018), SamSum (Gliwa et al., 2019)). This extensive manual evaluation of diverse systems and datasets aligns well with our task, and we adopt the ACU score as the gold score $S_{i,j}$ to predict. For our test set, we uniformly sampled 20% of the documents from each dataset in RoSE, totaling 500 documents. The remaining 2,000 documents were used for the training set. Data statistics can be found in Table 1.

Dataset	#Doc	#Sys.	#ACU
CNNDM	1,500	8-12	17.2k
XSum	500	8	2.3k
SamSum	500	8	2.3k

Table 1: Distribution of RoSE Dataset. Taken from Liu et al. (2023).

4.2 Preliminary Analysis - Do Systems Fail on The Same Documents?

In this section, we investigate whether different systems tend to consistently fail or succeed across the same documents. Since each system generates summaries for the same set of documents, their performance scores can be used to rank the documents. Our hypothesis is that different systems rank documents similarly, meaning the systems on certain documents consistently perform well or poorly across various systems.

To test this hypothesis, we measure the correlation between the ACU scores of summaries generated by different systems for the same document. Specifically, we calculate the correlation between the ACU scores of systems l and k as $corr(S_{l,1..N}, S_{k,1..N})$. We then compute the average correlation across all systems using the formula:

$$\frac{2}{M^2 - M} \sum_{l=1}^{M} \sum_{k=1}^{l-1} corr(S_{l,1..N}, S_{k,1..N}) \quad (2)$$

We obtained a Kendall Tau correlation of 0.446 and a Spearman correlation of 0.565 over the entire RoSE dataset, indicating a moderate agreement in document rankings across systems. This suggests that many documents retain their relative ranking, regardless of the system used for summarization. These findings support our assumption that certain document-specific features significantly influence summarization performance. As a result, it might

be possible to predict a document's average summarization performance based solely on its intrinsic characteristics.

5 Experiments

We examine the ability of several methods to rank the documents according to the average score across all systems, $d_j^* = \frac{1}{M}\sum_i^M S_{i,j}$. In this section we present simple baseline models based on document-based features (Section 5.1) and several supervised models trained on our training data (Section 5.2). In addition, we compare these methods to conventional summarization automatic metrics that utilize not only the document itself, but also the generated summary or even a reference summary (Section 5.3).

5.1 Baseline Models

Document Statistics. We explored several basic statistics about the documents, including document length by word, the number of numerical values, and the number of unique named entities identified using the NER module from the NLTK package. All of these features might be associated to the reading complexity of documents, where longer documents with more numeric details and name entities might be more complex to read.

Flesch–Kincaid Readability Tests. We applied the Flesch Reading Ease test (Flesch, 1948) to measure how easy the document is to read, with scores ranging from 1 to 100 where higher scores indicate easier readability. Similarly, we used the Flesch-Kincaid Grade Level to estimate the U.S. education grade level required to understand the text, with higher scores corresponding to a more advanced reading level.

5.2 Supervised Models

We trained several models with a fixed number of 5 epochs over our training data. More implementation details are elaborated in Appendix A.1.

Regression. In this model, we trained the model to predict the actual d_j^* scores. The input is a document j and the target is d_j^*. We leveraged the Longformer model (Beltagy et al., 2020) to allow a large input length of 4096 tokens required in many documents from our set. We added a regression head on top of the output of the 784 dimensional [CLS] token of the last layer of the Longformer. The regression head contains a feed forward component with an output of one dimension, which should predict the d_j^* score. Additionally, we used a standard MSE loss function for training. We denote this model as PreSummReg.

Classification. Since some applications may only care about document rankings rather than their exact d_j^* scores, we train an additional model to rank the documents directly based on pairwise comparisons. This approach aligns with our evaluation metric, which is based on correlation. Instead of predicting individual scores, we aim to determine which document in a pair should be ranked higher, and then aggregate these local decisions to form a global ranking.

In this model, the input consists of a pair of documents, and the task is to classify whether the first document should be ranked higher than the second. We represent each document using embeddings from a pre-trained Longformer model, concatenate the two embeddings, and feed the result into a RankNet model (Burges et al., 2005). The target label for each pair of documents (i, j) is $\delta_{ij} = \mathbb{1}_{d_i^* > d_j^*}$, where $\mathbb{1}$ is an indicator function that equals 1 if the condition $d_i^* > d_j^*$ is fulfilled, or 0 otherwise. The model is trained using a binary cross-entropy loss function.

We generate all possible n^2 pairs from the training set to fine-tune the model and use all m^2 pairs from the test set during evaluation. To derive the final global ranking, we follow the method outlined by Keswani and Jhamtani (2021), where the final score for document i is defined as $S(i) = \sum_j \hat{\delta} ij$, with $\hat{\delta} ij$ representing the predicted outcome for the document pair (i, j). We then sorted $S(i)$ scores for the final ranking. We denote this model as PreSummClas.

Frozen Weights. Given the relatively small amount of training data, we explored a model with fewer trainable parameters to better handle the limited dataset. Specifically, we employed a variant of the regression model, freezing all weights except those in the regression head. We denote this model as PreSummRegFroz.

5.3 Summarization Automatic Metrics

While our approach focuses on methods that rely solely on the document, we compare them to automatic summarization metrics that do not share this limitation. Specifically, we compare against two reference-free metrics, BLANC (Vasilyev et al., 2020a) and SummaQA (Scialom et al., 2019a),

Method	Kendall τ	Pearson r	Spearman
Document length	-0.005	-0.048	-0.010
Count of Numbers	-0.016	-0.106	-0.023
# Unique Named Entities	-0.054	-0.134	-0.071
Flesch Reading Ease	-0.016	0.030	-0.021
Flesch Kincaid Grade	-0.0489	-0.0483	-0.0104
PreSummReg (Ours)	**0.321**	**0.463**	**0.460**
PreSummClas (Ours)	0.306	0.389	0.389
PreSummRegFroz (Ours)	0.279	0.406	0.403
Blanc	0.246	0.322	0.347
SummaQA	0.167	0.241	0.286
ROUGE	0.431	0.652	0.597
BERTScore	0.252	0.430	0.351

Table 2: Test set correlations of different PreSumm methods. The top methods use the document only, while the bottom section of methods also uses the system summary.

which use both the document and the system-generated summary, as well as two reference-based metrics, ROUGE (Lin, 2004a) and BERTScore (Zhang et al., 2019), which evaluate the system summary against a reference summary. To obtain a per-document score using these metrics, we averaged the metric scores assigned to all summaries generated from the same document.

5.4 Results

All models are evaluated by measuring their correlation with the gold-standard ACU labels, which reflects how well they ranked the documents. The correlation scores are summarized in Table 2. Most baselines show near-zero correlation, except for the number of numeric values and named entities, which exhibit a small negative correlation. This suggests that an increased presence of numbers and entities may lead to lower summarization model performance. In contrast, most trained models achieve a moderate correlation, supporting our hypothesis that document ranking can, to some extent, be predicted based solely on the document content.

The table shows that our supervised models outperform all baseline models. Notably, they even surpass both reference-free metrics, despite these metrics utilizing the system summary, as well as BERTScore, which leverages both the system and reference summaries—whereas PreSummReg relies solely on the document. Although ROUGE outperforms our supervised models, it also benefits from access to both the system and reference summaries. These comparisons highlight the strong performance of our supervised models. Given its superior performance and simplicity, we selected the PreSummReg model to represent the PreSumm approach in the subsequent experiments.

Dataset	#Topics	#Sys.
DUC 2006	20	22
DUC 2007	23	13
TAC 2008	48	57
TAC 2009	44	55
TAC 2010	44	55
TAC 2011	44	50

Table 3: Number of topics (document sets) and system summaries that participated in Pyramid evaluation by dataset

5.5 Out-of-Distribution Performance

Our PreSummReg model was trained and evaluated on specific models and datasets. In this section, we investigate its ability to generalize to other datasets with systems not seen during training.

To that end, we adopted the Pyramid human annotations (Nenkova and Passonneau, 2004), which have been applied over several years to DUC and TAC multi-document summarization benchmarks, involving approximately 50 different systems per year (details in Table 3). To the best of our knowledge, this is the only resource for human evaluation of automatic summarization that does not involve the datasets we have used for training PreSummReg. Thus, it serves as an ideal benchmark to assess generalization.

We measured the correlation of PreSummReg to the Pyramid scores, and compared the performance against several baselines. To apply our single-document-based model and other baselines to the multi-document setup, we calculated scores for each document individually and then averaged them across each document set. As presented in Table 4, PreSummReg consistently outperforms the baselines across most datasets by a large margin, demonstrating its strong generalization capability to unseen models and datasets.

In addition to *system* performance, the Pyramid annotation method also evaluates the quality of *reference* summaries in some datasets. These datasets include document sets paired with four reference summaries. Pyramid evaluation is conducted on each reference summary by comparing it against the other three, with higher scores indicating greater alignment among the reference summaries and consistent content selection by different

Dataset	DUC 2006			DUC 2007			TAC 2008			TAC 2009			TAC 2010			TAC 2011		
	K	P	S	K	P	S	K	P	S	K	P	S	K	P	S	K	P	S
Length	**-0.26**	-0.34	**-0.39**	0.00	0.10	0.00	0.01	0.03	0.02	0.15	0.14	0.20	-0.02	-0.11	-0.02	-0.10	-0.07	-0.16
# Unique NEs	-0.25	-0.27	-0.37	0.06	0.27	0.12	-0.01	-0.01	0.01	**0.23**	0.25	**0.36**	0.07	0.08	0.13	0.16	0.25	0.19
Flesch	0.05	0.30	0.15	0.04	0.06	0.08	-0.03	-0.05	-0.03	0.18	0.24	0.25	0.03	-0.01	0.03	0.28	0.37	0.42
PreSummReg	0.25	**0.39**	0.37	**0.31**	**0.43**	**0.46**	**0.13**	**0.21**	**0.19**	0.13	**0.27**	0.21	**0.25**	**0.43**	**0.39**	**0.40**	**0.44**	**0.60**

Table 4: Correlation scores (Kendall-tau [K], Pearson [P], Spearman[S]) of various methods to different Pyramid dataset scores of *system* summaries

Dataset	TAC 2008			TAC 2009			TAC 2010			TAC 2011		
	K	P	S	K	P	S	K	P	S	K	P	S
Length	**-0.15**	-0.30	**-0.22**	-0.06	-0.05	-0.06	-0.07	-0.14	-0.11	-0.22	-0.17	-0.31
# Unique NEs	-0.14	**-0.31**	-0.21	-0.02	-0.07	-0.03	-0.05	-0.04	-0.07	-0.05	0.04	-0.07
Flesch	-0.05	-0.08	-0.10	0.03	**0.14**	0.06	0.02	0.12	0.03	0.23	0.33	0.31
PreSummReg	0.04	0.09	0.05	**0.09**	**0.14**	**0.12**	**0.15**	**0.23**	**0.22**	**0.32**	**0.47**	**0.46**

Table 5: Correlation scores (Kendall-tau [K], Pearson [P], Spearman[S]) of various methods to different Pyramid dataset scores of *reference* summaries

human summarizers.

This raises the question: Can we predict the agreement among *human summarizers* using only the source document? To explore this, we averaged the Pyramid scores of all reference summaries for each document set to obtain d_j^* and measured the correlation with PreSumm scores. As shown in Table 5, PreSumm outperforms all baselines in most datasets. This suggests that PreSumm can predict not only the performance of *models* but also the likelihood of *humans* agreeing on content selection using only the document. In other words, PreSumm can distinguish between documents with a clear main point that most people agree on and those lacking a central idea.

6 Extrinsic Evaluation through Downstream Tasks

In this section, we explore practical applications that benefit from predicting in advance the summarization model performance over documents. Additionally, these applications serve as extrinsic evaluations of our model. Specifically, we examine two use cases: (1) identifying in advance the documents where models perform poorly to enable manual summarization in hybrid systems, and (2) filtering out noisy documents in a multi-document summarization setting.

6.1 Selecting Documents for Manual Summarization in Hybrid Systems

Here, we focus on a use case within a hybrid summarization system, where a fixed percentage of documents can be manually summarized within the available budget. Accordingly, we aimed to assess whether PreSummReg can effectively identify, in advance, documents that models are likely to fail on, allowing these documents to be prioritized for manual summarization. The goal is to maximize the overall score of the entire document set.

For this experiment, we used the generated summaries of two systems from our test set, Pegasus (Zhang et al., 2020) and BART (Lewis et al., 2020). For evaluation, we used the average of the human ACU scores, where instead of manually summarizing a selected document, we assigned it an ACU score of 1. For the manual summarization budget, we selected either 10% or 20% of the test set documents. For statistical significance testing, we used the paired bootstrap test (Efron and Tibshirani, 1994) as explained in (Berg-Kirkpatrick et al., 2012). The detailed significance algorithm is provided in Algorithm 1 in the Appendix.

In addition to PreSummReg as a selection method, we evaluated several baseline methods. Baselines included Random selection and ranking methods based solely on the source document features, such as Flesch Reading Ease and the number of unique named entities. We also tested reference-free metrics such as Blanc (Vasilyev et al., 2020b) and SummQA (Scialom et al., 2019b). However, a limitation of these reference-free metrics is that they require system-generated summaries, unlike PreSummReg. For comparison, we included reference-based metrics such as ROUGE-2 F1 (Lin, 2004b), BERTScore F1 (Zhang et al., 2019), and

Based on	Selection Method	BART		Pegasus	
		Replace 10%	Replace 20%	Replace 10%	Replace 20%
Source	Random	0.425*	0.478*	0.394*	0.452*
	Num. Entities	0.437*	0.508*	0.416	0.486*
	Flesch	0.427*	0.494*	0.403*	0.472*
	PreSummReg	**0.451**	**0.525**	**0.423**	**0.499**
Source + Sys.	Blanc	0.435*	0.512*	0.416	0.496
	SummQa	0.434*	0.502*	0.414	0.481*
Sys.+ Ref.	Rouge	0.459	0.540	0.432	0.515
	Bert	0.454	0.527	0.429	0.507
	Meteor	0.459	0.538	0.433	0.516

Table 6: Averaged ACU scores of system summaries after replacing selected summaries with manual summaries. Scores significantly worse than PreSummReg are marked with *.

METEOR (Banerjee and Lavie, 2005), which represent an upper bound since except the system summary they rely on the reference summary as well—an unavailable resource in real-world scenarios.

As shown in Table 6, PreSummReg significantly outperforms all source-only and reference-free baselines in most cases, and approach the upper bound set by reference-based metrics. Overall, these experiments demonstrate that the PreSummReg model can effectively identify in advance documents that models are likely to fail on, optimizing the summarization process by saving time and resources.

6.2 Multi-Document Summarization

In a Multi-document summarization (MDS) task, a set of documents on the same topic needs to be summarized. However, these sets often include noisy documents that can negatively impact model performance (Giorgi et al., 2023). Additionally, summarizing a large number of documents poses challenges due to high input length, which can be costly or constrained by the token limits of certain models. Conventional MDS approaches typically concatenate all documents and truncate the input to meet the model's token limit or user budget, leading to the exclusion of some documents. This raises the question: can we achieve better results by using PreSumm to identify and exclude noisy documents from the set?

To test this, we adopted the MDS MultiNews dataset (Fabbri et al., 2019) and used the Pegasus and BART summarization models. We conducted summarization experiments with token limits of 256, 512, and 1024, where documents were truncated once the token limit was exceeded. In each experiment, we tested two variations: one where the documents were kept in their original order, and another where the documents were reordered based on their PreSummReg scores, with the lowest-scored documents placed at the end. It is important to note that reordering the documents has an additional benefit—better document sequencing can enhance summarization quality even without excluding documents, as suggested by Zhao et al. (2022). Thus, we aimed to determine whether combining PreSummReg-based document exclusion with PreSummReg-based reordering would lead to improved summarization outcomes.

# Tokens	Order	Pegasus			Bart		
		R-1	R-2	R-L	R-1	R-2	R-L
1024	Original	45.8	18.5	24.3	37.04	13.93	20.29
	PreSumm	**46.1**	**18.9**	**24.6**	**37.34**	**14.27**	**20.52**
512	Original	43.9	17.2	23.5	35.58	12.99	19.58
	PreSumm	**44.4**	**17.9**	**24.0**	**35.97**	**13.48**	**19.94**
256	Original	39.8	14.7	21.5	33.51	11.79	18.57
	PreSumm	**40.2**	**15.3**	**21.9**	**34.03**	**12.38**	**19.00**

Table 7: ROUGE scores for multi-document summaries generated using different document ordering methods, truncated after a specified number of tokens.

As shown in Table 7, the PreSumm-ordered documents consistently achieve higher ROUGE scores compared to the original document order across all input length limits. PreSumm is significantly better across all metrics according to the Wilcoxon Rank Test (Wilcoxon, 1945), except for 1024-token limit with R-1 and R-L. This demonstrates that using PreSummReg to identify independent documents that models are likely to summarize unsuccessfully is also effective in enhancing multi-document summarization settings, as such independent documents

contribute noise to the entire document set.

7 Analysis

In this section, we aim to investigate PreSummReg to better understand the properties that make a document less likely to be successfully summarized by various summarization systems. To that end, we examine the influence of document-based features over PreSummReg by measuring correlation (Section 7.1). Then, we conducted a manual analysis to reveal more insights and explain the automatic results (Section 7.2). Additional analysis examines how PreSumm deals with different corruptions can be found in Appendix B.

7.1 Document Feature Correlations

To gain deeper insights into which document-based features most strongly influence the performance of summarization systems over a certain document, we analyzed the correlations between various document characteristics and the PreSummReg score. For features, we leveraged the baseline methods from Section 5.1, including the document statistics and the reading ease tests. In addition, we added the salient sentence location feature that uses the reference summary, and therefore it could not be used as a method in Section 5.1.

Salient Sentence Location. Previous work pointed that the salient information in news documents tend to be at the beginning of the document (Lebanoff et al., 2019). Since in this work we use mostly news domain datasets, we hypothesize that models may struggle to generate effective summaries when the main theme appears in the middle of the text or when multiple themes are present within the document. Therefore, we would like to set the location of the salient sentences in the document as a feature. To determine the location of key sentences within a document, we adopted a method similar to (Nallapati et al., 2017; Chen and Bansal, 2018), where each sentence in the document is ranked by its similarity to the reference summary, using ROUGE scores as the metric. We selected the top-5 or top-10 most salient sentences, and their positions were normalized by the total number of sentences in the document. The average of these normalized indices represents the typical location of the most important sentences in the document.

The correlation results are shown in Table 8. The most strongly correlated feature is the location of salient information. The negative correlation indicates that the earlier the key information appears in the document, the easier it is for the model to summarize. This observation aligns with expectations, as this is a well-known characteristic of the news domain, where essential information is typically presented at the beginning.

Feature	Correlation
Document length	-0.0956
Count of Numbers	-0.0576
# of Unique Named Entities	-0.120
Flesch Reading Ease	0.182
Flesch Kincaid Grade	-0.166
Avg Loc of Salient Sent (top 10)	-0.266
Avg Loc of Salient Sent (top 5)	-0.305

Table 8: Pearson R correlations of document features to PreSummReg scores.

The Flesch Reading Ease and Flesch-Kincaid Grade Level scores show positive and negative correlations, respectively, indicating that more complex documents tend to result in poorer summarization model performance. In the same way, all basic document statistics show a relatively weak negative correlation, suggesting that greater document complexity—whether in length, the number of numerical values, or named entities—negatively impacts summarization performance.

Overall, while the correlation scores and their signs are consistent with our expectations, none of the features exhibited strong correlations. Consequently, we conducted a manual analysis to shed light on new document-based insights.

7.2 Manual Analysis

To gain deeper insights into document performance, we conducted a manual analysis. Specifically, one of the authors read the 15 best and worst-ranked documents according to PreSummReg predictions. In general, this analysis found three unique properties of the bottom-ranked documents.

Content Complexity. The main and most common characteristic of low-ranked documents is that they often cover complex topics, such as science or politics, which include numerous numbers, intricate details, and long, difficult-to-follow sentences and documents. In contrast, the top-ranked documents were typically much shorter (some with only a single sentence), with simple words and topics.

Coherence. Some bottom-ranked documents exhibited weak sentence-to-sentence connections or

lacked sufficient background information, starting abruptly in the middle of a story, counting on specific terms and knowledge of a specific unique field. Sometimes it happens because of the crawling process of documents, that includes the image caption or sub-headers in the middle of the text. Such crawling issues were also seen in the top-ranked documents, but less frequently.

Theme Change. Some low-ranked documents contain multiple, almost disjointed themes, making it difficult to determine which theme should be prioritized in the summary. This issue was especially pronounced in cases where the main theme was in the middle, requiring summarization models to go beyond its usual focus at the beginning of the document.

Overall, the low-ranked documents were notably more difficult to read, often requiring re-reading of certain sentences for comprehension, whereas the top-ranked documents were much more fluent and easier to follow. Examples of documents with these challenges are provided in Appendix D.

To validate these findings, we conducted an extended annotation study with five NLP research students. They annotated a total of 50 documents (10% of the test set), with each document being annotated by two students. Each annotation task involved a pair of documents: one from the 25 lowest-ranked by PreSummReg and another randomly selected document. Following the insights from our preliminary manual analysis, annotators first identified independent flaws in each document, such as coherence issues or theme shifts. Then, they performed a comparative evaluation, determining which document contained more complex content and which was more fluent overall.

Due to the subjective nature of this task, there were disagreements over many annotation instances. To resolve these, we conducted a second round in which annotators discussed the disputed cases, attempting to reach a consensus. After this process, the inter-annotator agreement, measured by Cohen's Kappa coefficient (Cohen, 1960), reached 0.76 for Coherence, 0.66 for Theme-change, 0.92 for Complexity, and 0.53 for Fluency.

As can be seen in Table 9, our expanded analysis confirms the initial observation. Documents with lower PreSummReg scores are more prone to coherence issues and theme changes, are more complex, and are less fluent. This confirms that documents difficult for humans to read tend to be more challenging for summarization models as well.

	Independent		Comparable	
Doc Type	Coherence	Theme	Complex	Fluent
Random	28%(44%)	12%(12%)	32%(36%)	60%(84%)
Low PreSumm	68%(76%)	20%(44%)	64%(68%)	16%(40%)

Table 9: Percentage of cases where all annotators (*at least one annotator*) identified an *Independent* flaw in a document or preferred one document on *Comparable* criteria, categorized by document type.[2]

These results also align with the correlation scores presented in Section 7.1. As noted, Reading Ease tests can indicate complexity, while the placement of salient information may relate to theme-changes and overall fluency. Although the correlation scores for these features were relatively weak compared to our manual analysis, this discrepancy may arise because our manual analysis focused on the lowest-scored documents rather than the full dataset. While this subset is more practically relevant, it also amplifies these properties, potentially leading to a stronger observed signal.

To compare this analysis to the actual performance rather than predicted by PreSummReg, we further examined documents with low average gold ACU scores. Interestingly, similar patterns emerged. However, some of these documents were penalized in ACU scoring due to misalignment between the reference summary and document content, rather than inherent flaws in the document itself. Since our model does not rely on reference summaries, it was unaffected by this misalignment and thus did not always rank these documents as low-quality.

8 Conclusion

In this work, we introduced PreSumm, a novel approach that opens up new research avenues in understanding the structural features that make a document less likely to be summarized successfully. Our findings suggest that documents that are more complex to read for humans are also ranked low by PreSumm models, implying the centrality of this feature for summarization models. We hope these insights will contribute to the design of more robust and effective models in the future.

[2]The differences in percentages between the low PreSummReg documents and the random set are statistically significant in all cases according to paired bootstrap testing (Efron and Tibshirani, 1994), except for Theme Change with full annotator agreement.

Limitations

Our study covers the RoSE dataset extensively and focuses on summarization of the news domain only. Therefore, we cannot explore the complete space of summarization systems and we are limited to both the datasets and summarization systems that RoSE provides due to our extensive use of the manual ACU score. Because of this, some of the results could be due to other factors that relate to the dataset and would not generalize strongly outside of this study or to other domains.

Although we showed in our work that our model works relatively well on out-of-distribution data, we did not examine dataset out of the news domain. Therefore our conclusions are limited to this domain.

In future works, we would seek to train a similar model on a larger dataset. However, using ACU scores would be difficult because of the human labor involved, which could be avoided by using an annotated metric to train on. However, will make us biased towards the chosen metric, which is another limitation.

Out of distribution data is a major factor when it comes to model performance. To mitigate this would require greatly increasing the scope of the experiment and to train on a broader dataset for more accurate predictions.

Overall, overcoming these limitations would necessitate a much larger corpus with either a large set of automated or human annotated metrics to perform a similar study on a much larger set of the space of documents and summarization system.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. We would like to thank Lorenzo Flores, Sienna Hsu, and Cesare Spinoso-Di Piano for their annotations. This work was supported in part by the IVADO Postdoctoral Fellowship, Canada CIFAR AI Chair, and the Natural Sciences and Engineering Research Council of Canada.

References

Albaraa Abuobieda, Naomie Salim, Ameer Tawfik Albaham, Ahmed Hamza Osman, and Yogan Jaya Kumar. 2012. Text summarization features selection method using pseudo genetic-based model. *2012 International Conference on Information Retrieval & Knowledge Management*, pages 193–197.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Araly Barrera and Rakesh Verma. 2012. Combining syntax and semantics for automatic extractive single-document summarization. In *Computational Linguistics and Intelligent Text Processing*, pages 366–377, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2022. PreQuEL: Quality estimation of machine translation outputs in advance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11183, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Mohamed Abdel Fattah and Fuji Ren. 2009. Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. *Computer Speech & Language*, 23(1):126–144.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Wang, and Arman Cohan. 2023. Open domain multi-document summarization: A comprehensive study of model brittleness under retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8177–8199, Singapore. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv*, abs/2209.12356.

Pankaj Gupta, Vijay Shankar Pendluri, and Ishant Vats. 2011. Summarizing text by ranking text units according to shallow linguistic features. *13th International Conference on Advanced Communication Technology (ICACT2011)*, pages 1620–1625.

Vishal Keswani and Harsh Jhamtani. 2021. Formulating neural sentence ordering as the asymmetric traveling salesman problem. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 128–139, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Uday Kulkarni and Rajesh Shardanand Prasad. 2010. Implementation and evaluation of evolutionary connectionist approaches to automated text summarization. *Journal of Computer Science*, 6(11):1366–1376.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring Sentence Singletons and Pairs for Abstractive Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

Chin-Yew Lin. 2004a. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156, Tokyo, Japan. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Rajesh Shardanand Prasad, Nitish Milind Uplavikar, Sanket Shantilals Wakhare, VY Jain, Tejas Avinash,

et al. 2012. Feature based text summarization. *International journal of advances in computing and information researches*, 1(2):15–18.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019a. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019b. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020a. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020b. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468, Seattle, United States. Association for Computational Linguistics.

Frank. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics*, 1:196–202.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Chao Zhao, Tenghao Huang, Somnath Basu Roy Chowdhury, Muthu Kumar Chandrasekaran, Kathleen McKeown, and Snigdha Chaturvedi. 2022. Read top news first: A document reordering approach for multi-document news summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 613–621, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

The SummEval models were built using the Longformer architecture, each with distinct output heads and training methodologies tailored to the task.

All models were trained for 5 epochs with a learning rate of 1e-6, using an 80%-20% train-validation split of the dataset. The base model used was the "allenai/longformer-base-4096" (Beltagy et al., 2020), configured to handle the maximum sequence length.

A.1.1 Regression Head

For regression tasks, a single linear layer was added to map the 768-dimensional CLS token embedding to a one-dimensional output, representing the predicted score.

A.1.2 Classification Head

For classification tasks, the Longformer backbone was paired with a more complex classification head. This head comprised a feedforward neural network with the following structure: a linear layer mapping 768 dimensions to 512, a ReLU activation, and a second linear layer reducing 512 dimensions to a single scalar output. During the forward pass, the model computed individual scores from two heads, denoted as $s1$ and $s2$, and generated the final probability by subtracting these scores and applying the sigmoid function.

B Document-Based Transformation Analysis

To further investigate the influence of document-based features on model performance, we explored how the predicted score of a document changes when specific features are perturbed. By comparing the predicted scores before and after these transformations, we can gauge the importance PreSumm places on each feature. Intuitively, the transformations causing the largest change in predicted scores should indicate which features have the greatest impact on document performance in summarization. We applied these transformations to our test set. Below, we detail the transformations applied.

Removing content. We tested several removal strategies: removing the first sentence (often critical for summarization), removing 5 or 10 salient sentences (as defined in Section 7.1), and randomly removing 30% of the words or

Transformation	Src.	Trans.	Delta
Remove 10 salient sentences	0.528	0.428	-0.100
Keep last 3 sentences	0.528	0.441	-0.0877
Delete 30% of words	0.528	0.470	-0.0584
Remove 5 salient sentences	0.528	0.476	-0.0523
Randomly Shuffle of Sentences	0.528	0.486	-0.0427
Move 10 salient sentences to end	0.528	0.502	-0.0269
Keep first 3 sentences	0.528	0.553	0.0246
Remove first sentence	0.528	0.506	-0.0225
Move 5 salient sentences to end	0.528	0.509	-0.0199
Replace names w/ from bank	0.528	0.512	-0.0163
Replace names w/ spacy name	0.528	0.518	-0.0102
Corrupt Grammar	0.528	0.520	-0.008
Append contradictions	0.528	0.533	0.005
Delete 30% of sentences	0.528	0.530	0.00150

Table 10: Predicted PreSummReg scores of documents before a transformation (Src.) and after (Trans.)

sentences to disrupt fluency and coherence. Additionally, we removed all sentences except the first three or last three to assess the importance of content location.

Moving content. Given that salient information location was identified as a key feature in Section 7.1, we moved the salient sentences to the end of the document. We also randomly shuffled all sentences to disrupt coherence.

Replacing content. We replaced named entities to test the impact on consistency. We also introduced contradictions by adding negation sentences and corrupted the grammar by converting all verbs to their lemma forms.

As expected, the most impactful corruptions, with the highest changes in predicted scores, were removing the 10 most salient sentences and removing all content except for the last three sentences. Other significant transformations included deleting 30% of the words, removing 5 salient sentences, and shuffling sentences randomly.

Surprisingly, moving the salient sentences to the end of the document had little effect, despite the location of salient information being one of the most influential features in Section 7.1.

Interestingly, content replacement, such as grammar corruption or adding contradictions, did not significantly affect PreSumm's performance. It also appears that strong perturbations, such as deleting 30% of sentences, did not lead to large differences in scores. This might be because these artificial corruptions are not natural and therefore deviate too much from the patterns seen during model training, causing the model to mispredict their impact on summarization performance.

Algorithm 1 PreSumm vs method X Paired Bootstrap Significance Test

Input: Test set of documents and system summaries of a specific system, PreSumm scores, X scores, ACU scores
Output: p-value
Extract the current difference in performance between PreSumm filtering and X filtering, denoted as `original_diff`.
Initialize $s = 0, b = 10,000$
n is the number of instances (documents)
for $i \leftarrow 1$ to b **do**
    Sample n instances with replacement from the test set
    Replace 10% or 20% of system summaries with reference summaries according to PreSumm, and average the ACU scores to get `PreSumm_filter_score`
    Replace 10% or 20% of system summaries with reference summaries according to X, and average the ACU scores to get `X_filter_score`
    if PreSumm_filter_score - X_filter_score > $2 \times$ `original_diff` **then**
        $s = s + 1$
    end if
end for
Compute $p\_val = \frac{s}{b}$

C Annotation Guidelines

Below are the annotation guidelines provided to annotators for the manual analysis described in Section 7.2.

Read the following documents carefully and answer the questions below for each.

Document A:
<document A>

- Did you notice any coherence issues in Document A? Examples: corruptions, unrelated sentences, topics starting abruptly without enough background.

  Yes/No

- Does Document A contain a theme change? Examples: discusses topics that deviate significantly from the main theme, or there are multiple main themes.

Yes/No

Document B:
<document B>

- Did you notice any coherence issues in Document B? Examples: corruptions, unrelated sentences, topics starting abruptly without enough background.

  Yes/No

- Does Document B contain a theme change? Examples: discusses topics that deviate significantly from the main theme, or there are multiple main themes.

  Yes/No

Comparison Questions:

- Which document had more complex content?

  Document A/Document B

- Which document is more fluent to read?

  Document A/Document B

D Example Documents

Tables 11 and 12 present examples of documents from the 15 lowest PreSummReg scores, annotated with their respective challenges. For comparison, Table 13 provides examples of documents from the 15 highest PreSummReg scores.

Challenging Characteristics	Document
Coherence	Judge Thokozile Masipa did the same for the lawyers on Thursday, urging them to make good use of the upcoming fortnight break for the Easter holidays. In that spirit, here are a few questions that have been niggling me in recent days. Tweet your thoughts and suggestions to @BBCAndrewH. I will be taking a week off and then focusing on South Africa's general election before returning to the hard benches of Courtroom GD on 5 May.
Theme Change	The images, taken by Syd Shelton, from Pontefract, include pictures of The Clash, Misty in Roots and The Specials. The collection also features photos taken at the Rock Against Racism Carnival at Victoria Park, Hackney, which attracted a crowd of 100,000. The show runs from Friday to 3 September at the Impression Gallery. The Rock Against Racism (RAR) movement formed in response to controversial remarks made by Eric Clapton in 1976. In the following years, RAR staged marches, festivals and more than 500 concerts in the UK in a bid to fight racism through music. Shelton, who studied Fine Art in Leeds and Wakefield, said he became involved with the movement after returning to the UK from America in 1976. He said: Ï was appalled at the state of race relations in Britain, in particular things like the Black and White Minstrel Show and the signs I saw in some windows saying 'No Blacks, No Dogs, No Irish'. Ït was a pretty serious situation and I always loved music and very quickly hooked up with the people that had set up RAR. Ït was a bizarre mixture of people, photographers, graphic designers, writers, actors and, of course, musicians. Ẅe were very lucky in the sense that we tuned in to that explosion of punk and UK reggae and brought the two together. That said more about what RAR was about than any of the slogans we may have shouted from the stage.¨ He added: Ï hope the exhibition shows that you can change things and you can actually take a stand, even in the most difficult of situations. ...

Table 11: Example documents from the 15 lowest predicted scores by PreSummReg, categorized by their challenging characteristics.

Challenging Characteristics	Document
Content Complexity	Welsh language minister Alun Davies told AMs it would help efforts to reach that goal stay on the right track. Targets to meet growing demand for Welsh-speaking teachers and public sector workers will also be set. Culture committee chairwoman Bethan Jenkins said AMs had been told 70% more Welsh-medium teachers were needed. Mr Davies responded that around a third of teachers in Wales could speak Welsh, and that the challenge was to see if more of them would be willing to teach through the medium of Welsh. Earlier this month, Welsh language commissioner Meri Huws called for r̈adical changeïn the education system to ensure all children under the age of seven were ïmmersedïn Welsh.":
Content Complexity	He said new forests would slow flooding by trapping water with their roots. The idea of r̈ewildingẗhe uplands is catching on fast as parts of Britain face repeated flooding, with more rainfall on the way. Environment Secretary Owen Paterson said he would seriously consider innovative solutions like rewilding. The government has been criticised for being slow to capitalise on the benefits of capturing rain where it falls. Lord Rooker, a Labour peer, said too much emphasis had been attached to the look of the countryside rather than practical considerations like trapping water. Ẅe pay the farmers to grub up the trees and hedges; we pay them to plant the hills with pretty grass and sheep to maintain the chocolate box image, and then wonder why we've got floods,ẗhe said. The idea of reintroducing forests into catchments has been strongly supported by several leading scientists. The government is sponsoring a handful of catchment trials to assess the potential of the upstream areas to catch water and send it slowly downhill. ...

Table 12: Example documents from the 15 lowest predicted scores by PreSummReg, categorized by their challenging characteristics.

Document
Eve: Where are we meeting? Charlie: at the entrance Nicole: yes, it's the best place. We would't find each other inside, it'll be too crowded Eve: ok!
Jair: Still busy? Callum: Yes a little sorry Jair: ok
A 16-year-old girl is anxiously awaiting blood test results after sitting on a needle on a bus. Francesca Palmer-Norris was on the top deck of the number 24 Brighton and Hove Bus Company vehicle when she was pricked by the needle. The worried student, from Brighton, East Sussex spent the next four hours in hospital where she was given a hepatitis jab and had blood tests. Worried:Ŏ0a0Francesca Palmer-Norris is awaiting blood test results after sitting on a needle on the top deck of the bus . Speaking about the incident, Ms Palmer-Norris said: 'My friend and I had got on the bus to go home and we were sat on the top. 'I suddenly had this shooting pain in the back of my leg. I reached down and pulled out a needle that had snapped in half. 'Then I looked down the side of the bus seat and there were packets and a syringe on the floor and the rest of the needle.' She added: 'When the bus reached the next stop, I explained to the driver what had happened and he said it was best to go to the hospital.' She was given a jab and had blood tests before going home that night. Francesca Palmer-Norris was on the top deck of the number 24 Brighton and Hove Bus Company vehicle when she was pricked by the needle (stock image) Ms Palmer-Norris said: 'The worrying thing now is I am waiting for the results to come back. 'My head is all over the place - I can't sleep.' The bus company said the driver closed the top deck of the bus after the incident and took the vehicle for a full inspection as 'soon as practicably possible'. Adrian Tullett, head of operations at Brighton and Hove Bus Company, said the incident was being investigated using CCTV footage. He added: 'The driver followed procedure and secured off the top deck as soon as he was made aware of an object that needed removing from the seating area. 'He took the vehicle out of service for a full inspection as soon as was practically possible. 'We would like to reassure passengers we take these matters very seriously and that all our buses get a visual inspection at the end of each journey. Our customer services team is liaising direct with the girl's family.' Sussex Police is also investigating the incident.

Table 13: Example documents from the 15 highest predicted scores by PreSummReg.