LongDocFACTScore: Evaluating the Factuality of Long Document Abstractive Summarisation

Jennifer A Bishop[1], **Qianqian Xie**[1], **Sophia Ananiadou**[1,2]
[1]National Centre for Text Mining, Department of Computer Science,
The University of Manchester, United Kingdom
[2]Artificial Intelligence Research Center, Tokyo, Japan
{jabishop.research, xqq.sincere}@gmail.com, sophia.ananiadou@manchester.ac.uk

Abstract

Maintaining factual consistency is a critical issue in abstractive text summarisation, however, it cannot be assessed by traditional automatic metrics used for evaluating text summarisation, such as ROUGE scoring. Recent efforts have been devoted to developing improved metrics for measuring factual consistency using pre-trained language models, but these metrics have restrictive token limits, and are therefore not suitable for evaluating long document text summarisation. Moreover, there is limited research and resources available for evaluating whether existing automatic evaluation metrics are fit for purpose when applied in long document settings. In this work, we evaluate the efficacy of automatic metrics for assessing the factual consistency of long document text summarisation. We create a human-annotated data set for evaluating automatic factuality metrics, LongSciVerify, which contains fine-grained factual consistency annotations for long document summaries from the scientific domain. We also propose a new evaluation framework, LongDocFACTScore, which is suitable for evaluating long document summarisation. This framework allows metrics to be efficiently extended to any length document and outperforms existing state-of-the-art metrics in its ability to correlate with human measures of factuality when used to evaluate long document summarisation data sets. We make our code and LongSciVerify data set publicly available: https://github.com/jbshp/LongDocFACTScore.

Keywords: Evaluation Methodologies, Summarisation, Natural Language Generation

1. Introduction

Factual inconsistency, i.e., when a generated summary is not entailed by its source document, is a well-documented limitation of modern neural summarisation methods (Maynez et al., 2020; Wallace et al., 2021). Although Large Language Models (LLMs) have shown greatly superior performance on a range of NLP tasks, including summarisation (Zhang et al., 2023; Xie et al., 2023), and are increasingly being used for summarisation of long documents in real world applications, even the best performing models, such as GPT-4 (OpenAI, 2023), are flawed in their ability to remain factual consistent (Bang et al., 2023; Ye et al., 2023; Min et al., 2023).

Human evaluation is generally regarded as the gold standard for evaluating generative models, yet it is timely and costly to conduct, particularly for tasks involving long documents, and thus only a small proportion of long document summarisation studies perform a human evaluation on long document data sets (Krishna et al., 2023). Consequently, there is a requirement for effective automatic evaluation metrics which align to human judgement in long document settings.

Although ROUGE scoring (Lin, 2004) is the traditional metric for automatic evaluation of text summarisation, it is flawed and does not correlate well with human judgement (Yuan et al., 2021; Huang et al., 2020; Kryscinski et al., 2019). There have been efforts to develop improved model-based metrics for measuring factual consistency (Scialom et al., 2021; Yuan et al., 2021; Kryscinski et al., 2020; Qin et al., 2022; Fu et al., 2023; Liu et al., 2023), however, the studies proposing these metrics only conduct evaluation on short document summarisation data sets (Hermann et al., 2015; Grusky et al., 2018; Narayan et al., 2018; Pagnoni et al., 2021) and there is limited research or available data sets for evaluating these automatic metrics in long document settings.

Since modern evaluation metrics use pre-trained language models (PLMs), they are only able to process a limited number of tokens at a time and must truncate, on average, over half of the tokens of a long source document in their calculations. Therefore, they cannot be applied effectively when used in long document settings (Koh et al., 2022). This issue is exacerbated when evaluating factual consistency, where many of the metrics are designed to be reference-free (Yuan et al., 2021; Fu et al., 2023; Liu et al., 2023; Scialom et al., 2021; Kryscinski et al., 2020), i.e., they use the source document (generally a much longer document), rather than a gold summary, in their calculations.

In this work, we propose a reference-free evaluation framework, LongDocFACTScore, intended for assessing the factual consistency of abstractive summarisation of long documents. We show

LREC-COLING 2024, pages 10777–10789

that this framework outperforms all other automatic metrics evaluated in their correlation with human annotations of factuality on the long document data sets in our experiments.

Our proposed framework can be efficiently be extended to any length document and incorporates fine-grained, sentence-level assessments of factuality consistency to give a document-level score for the factual consistency of a summary. We conduct an evaluation of the efficacy and efficiency of LongDocFACTScore and other automatic evaluation metrics on a range of long and short document data sets.

Addressing the scarcity of resources for evaluating automatic metrics in long document settings, we create a long document data set of the scientific domain with fine-grained, expert, human annotations of factual consistency, which we make available alongside our code. We hope that this resource encourages future work into the evaluation of automatic metrics in long document settings.

2. Related Work

2.1. Automatic Evaluation Metrics for Evaluating Factual Consistency

ROUGE scoring (Lin, 2004), which uses word overlap between two texts to calculate their similarity, has long been the popular automatic metric used for evaluation of text summarisation. However, more recently, model-based metrics, such as BERTScore (Zhang et al., 2020b), which measures agreement at a token level between the cosine similarity of BERT-based (Devlin et al., 2019) embeddings, have shown improved correlation with human judgement. Additionally, reference-free model-based metrics have shown improved performance for the evaluation of factual consistency on short document summarisation data sets. FactCC (Kryscinski et al., 2020) uses a fine-tuned BERT-based classifier to predict, for each sentence of a summary, whether it is correct or incorrect, given its source document. QuestEval (Scialom et al., 2021) uses T5-based models (Raffel et al., 2020) for a question generation and answering approach. BARTScore (Yuan et al., 2021) uses BART (Lewis et al., 2020) to calculate the log probability of generating a sequence of text, given a second sequence, to predict an automatic score. T5SCORE (Qin et al., 2022) uses T5-based models and combines the generative approach taken by BARTScore with a discriminative approach - i.e., fine-tuning a model to predict a quality score. Unfortunately, many of these model-based metrics are costly to run, for example, QAGS (Wang et al., 2020), another question-answering based metric, running on a single NVIDIA v100 GPU, will take 4 days to process the CNN/DM test data set (Nan et al., 2021). Other recent works have proposed the use of LLMs for evaluation of NLP tasks (Fu et al., 2023; Luo et al., 2023; Liu et al., 2023), but still limit their evaluation to short document summarisation data sets.

2.2. Frameworks for Evaluation of Long Document Summarisation

There has been limited research into automatic evaluation metrics for long document summarisation, despite the value of summarisation being derived mostly when applied to long documents. Koh et al., (2022) carried out a survey of long document summarisation and found a gap in research for automatic metrics which could efficiently and effectively be applied to long document data sets. Krishna et al., (2023) propose guidelines for evaluation of long document summarisation, and provide LongEval, the only publicly available long document summarisation data set with human annotations of factual consistency which we could find at the time of conducting our research. SMART (Amplayo et al., 2022) proposes a method of extending evaluation metrics to long documents by cycling through them, but do not propose an efficient way of doing so, nor do they evaluate their work on long document summarisation data sets due to the lack of availability of these resources.

3. Methods

In this section, we describe the LongDocFACTScore framework. This framework builds on existing evaluation metrics but applies them in a novel way, providing a summary-level factuality score that considers fine-grained statements, whilst scaling efficiently though a document of any length. LongDocFACTScore evaluates each sentence in a predicted summary against the most similar sections of a source document, calculated using the cosine similarity of their sentence embeddings (Reimers and Gurevych, 2019). Individual summary sentences are evaluated to ensure fine-grained statements are considered in the predicted summary-level score, whilst sentence embeddings are used to improve efficiency of the framework.

To calculate LongDocFACTScore, both the source document $D = \langle s_i, i \in I \rangle$ and its generated summary $S = \langle s_j, j \in J \rangle$ are split into sentences using the nltk library[1]. Splitting a document into sentences before applying an evaluation metric has shown to be effective in prior works (Min et al., 2023; Amplayo et al., 2022). For each of these sentences, sentence embeddings are generated

[1] https://www.nltk.org

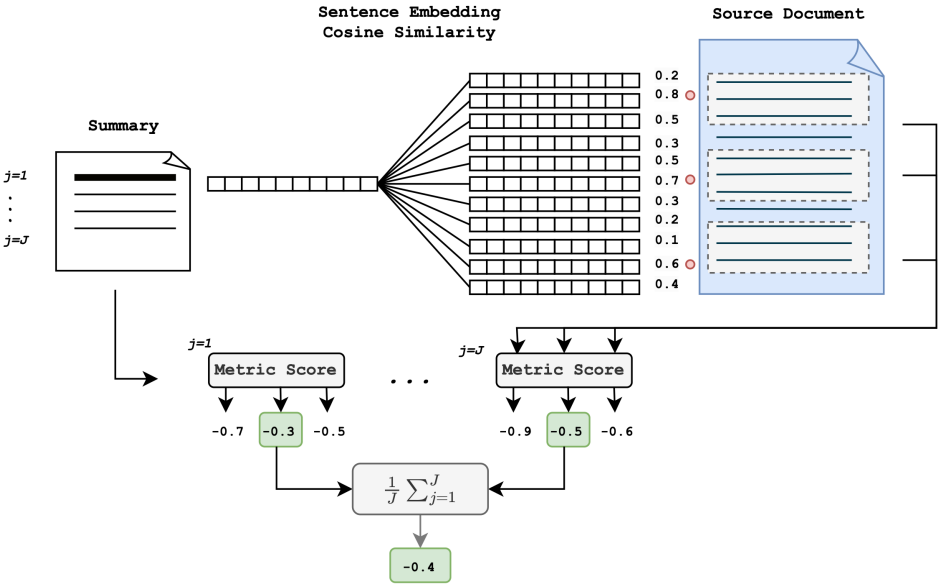

Figure 1: Illustration of the LongDocFACTScore framework.

using the sentence-transformers library[2] initialised with the bert-base-nmli-mean-tokens model[3]. For each sentence in the predicted summary s_j, the cosine similarity between its sentence embedding and the sentence embedding of each sentence in the source document s_i is calculated. D is then re-indexed by the cosine similarity scores, so that the new index k is sorted by:

$$arg\max_{i \in I} \left(cosine\_similarity\left(s_j, s_i\right)\right). \quad (1)$$

The K most similar source document sentences are then selected and are each concatenated with their preceding and following sentences, thus giving $s_k^* = s_{k-1} + s_k + s_{k+1}$, to create the sequence of slightly longer text snippets. We select the K most similar source document text snippets to improve the efficiency of the framework. We assume the most similar source document text snippets are the ones most likely to be relevant to make an assessment of factual consistency.

The metric score is then calculated between each of the source document text snippets s_k^* and the summary sentence s_j, and the maximum of these scores is taken. In this work, we set $K = 3$, a decision which we justify in Section 5.5.

For each sentence, s_j in S, of the generated summary, the process is repeated, resulting in one score per generated summary sentence. The mean of these scores is then calculated, providing an overall summary score given by the equation:

$$\frac{1}{J}\sum_{j=1}^{J} \max_{k=\{1,2,3\}} (metric(s_j \,|s_k^*)). \quad (2)$$

Figure 1 illustrates this framework, showing for a single sentence in the generated summary, the similarity scores being calculated for every sentence in the source document, and the resulting three highest scoring sentences being concatenated with their surrounding sentences. A metric score is then calculated between these three source document text snippets and the summary sentence. As indicated in Figure 1, this process is repeated for every sentence in the generated summary and the scores are averaged. For contrast, Figure 2 shows the method for directly applying an automatic scoring metric to a long document, without applying the LongDocFACTScore framework. The entire generated summary and the truncated long source document are directly input to the metric, resulting in one score. Consequently, there are three fundamental differences between LongDocFACTScore and an automatic metric designed for short document evaluation:

- The first difference is that LongDocFACTScore considers sections of text from the full length of the source document in its calculation (using sentence embeddings to select the most relevant from across the document) whereas other metrics truncate the source document. For metrics applied without the LongDocFACTScore framework, if a generated summary includes content from the latter part of

[2] https://github.com/UKPLab/sentence-transformers

[3] https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

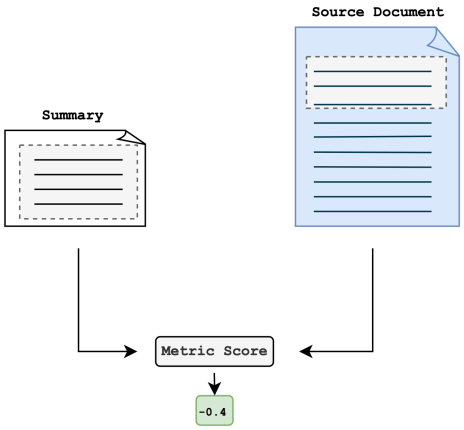

Figure 2: Calculation of a traditional automatic metric for assessing factual consistency.

a long document, it will be ignored, which is a problem when assessing factual consistency of long document summarisation.

- The second significant difference is that LongDocFACTScore calculates a metric score on short sections of text at one time, comparing one sentence in the predicted summary to a short section of the source document, rather than long, truncated sections. This allows for better evaluation of fine-grained statements.
- Lastly, LongDocFACTScore uses sentence embeddings to identify the most similar parts of the source document. This improves the efficiency of the framework, as it avoids the metric needing to be applied for each pair-wise set of sentences in the source document and summary.

4. Experimental Data Sets

We evaluate the automatic metrics on their ability to assess factual consistency on two long document data sets and several short document data sets. We collected our own long document data set, consisting of documents from the biomedical and scientific domains annotated by six expert human annotators with fine-grained factual consistency labels. We refer to this data set as the LongSciVerify data set and provide further details of its curation in Section 4.1. The data set is made available alongside our code. We further evaluate our methods on the LongEval PubMed data set (Krishna et al., 2023), another long document data set with factual consistency annotations. Finally, we conduct an evaluation on a range of short document data sets with human annotations of factuality, which have been used to evaluate automatic metrics in prior works (Yuan et al., 2021).

	Doc. tokens	Doc. sentences	Sum. tokens	Sum. sentences
PM	3209	124	208	9
AX	6515	249	279	11

Table 1: Average number of tokens and sentences in the evaluated data sets. PM denotes the PubMed data set and AX denotes the ArXiv data set.

4.1. The LongSciVerify Data Set

To support the evaluation of factuality metrics for long documents, we create a new data set called LongSciVerify, with multiple summaries generated from long documents, and fine-grained human annotation scores of their factual correctness. This data set consists of 270 annotated summaries generated from the long document, English-language PubMed and ArXiv data sets (Cohan et al., 2018). A description of the PubMed and ArXiv data sets can be found in Table 1.

From each of the PubMed and ArXiv data sets, fifteen articles were randomly sampled. Summaries were generated for these data sets using three different abstractive methods, which were all able to consider the entire long document in the generation of their summaries. These methods were selected to enable an effective evaluation of the performance of the automatic metrics in long document settings. Details of the abstractive methods used to generate the summaries are provided in Appendix A. As the PubMed and ArXiv data sets included in this data set are highly domain specific, we recruited six expert annotators, three per data set, to review the automatically generated summaries. At the time of evaluation, all of the expert annotators reviewing the PubMed data set were, or were in the final years of study to be, qualified clinicians. The expert annotators for the ArXiv data set had all achieved a minimum of an undergraduate degree in a physical science. The annotators who participated in our study were colleagues of the authors and therefore volunteered to participate in the study without payment. It was made clear to the annotators that this human evaluation was for scientific research on abstractive summarisation with the intention for use in a scientific publication.

The definition of factual consistency we provided to annotators was taken from Fabbri et al., (2021): *"Factual consistency: The factual alignment between the summary and the summarised source. A factually consistent summary contains only statements that are entailed by the source document."*.

We opted to capture a fine-grained binary classification metric (entailed vs not entailed), due to this having been shown to be effective and achieve higher inter-annotator agreement scores (IAA) in prior work (Krishna et al., 2023; Min et al., 2023).

	LSV ArXiv	LSV PubMed	LE PubMed
Fine-grained	0.54	0.76	-
Summary-level	0.70	0.82	0.61

Table 2: IAA of the human-annotated data for the LongSciVerify (LSV) and LongEval (LE) data sets, calculated using Krippendorf's alpha metric, for fine-grained and summary-level annotations of factual consistency.

Annotators were asked to mark a sentence as 'not entailed' if there were any factual inconsistencies. For each generated summary included in the study, we sampled three summary sentences and selected the most similar two text snippets (1-3 sentences) from the source document, calculated using sentence embeddings and cosine similarity. The human annotators were then given the three sentences sampled from the generated summary, and the corresponding two text snippets for each, from the source document and were asked to decide whether, given the text snippets, if each sentence was entailed or not. We provide an example screenshot of the factuality scoring for the three sampled sentences from a PubMed article summary in Figure 3.

For each of the PubMed and ArXiv samples, each of the three human annotators evaluated the same three summaries (generated by the three different methods) from the same 15 randomly sampled documents, thus resulting in 270 annotated summaries, with three fine-grained annotations per summary. During the human evaluation study, the annotators were unaware of which method was used to create each summary.

Table 2 shows the IAA of the fine-grained human annotated data, for each data set, calculated using the Krippendorff's alpha metric[4] (Krippendorff, 2004). In Table 2, the IAA is calculated both between fine-grained, sentence-level entailment annotations, and for the summary-level annotations (i.e., the average of the fine-grained annotations per summary). For our LongSciVerify PubMed data set, the IAA of the fine-grained factual consistency annotations is relatively high. However, the IAA of the LongSciVerify ArXiv data set is a little lower. We hypothesise this could be due to the noise in the ArXiv data set (Koh et al., 2022) and the highly domain-specific nature of the data set.

4.2. The LongEval Data Set

We additionally evaluated LongDocFACTScore and the other automatic metrics included in our study on the publicly available long document PubMed LongEval data set (Krishna et al., 2023). This data set consists of summaries generated from two abstractive models: LongT5-large (Guo et al., 2022) and BigBird-PEGASUS (Zaheer et al., 2020). Three expert annotators were hired to give annotations of factuality on 40 summaries (two summaries generated by different methods for 20 documents), giving 120 annotated summaries. The IAA of the summary-level human annotations of factual consistency is given in Table 2. As for the LongSciVerify data set that we create, the LongEval data set was created by averaging fine-grained annotations to give a summary-level factuality score. However, in constrast to the LongSciVerify data set, annotators considered the entire source article, rather than just the most relevant sections of it, when making their assessments.

5. Experiments

5.1. Experimental Setting

As baselines, ROUGE[5] and BERTScore were implemented. We additionally implemented SOTA reference-free metrics: FactCC[6], QuestEval[7], and BARTScore[8] (using the 'bart-large' model[9]). In this work, we apply the LongDocFACTScore framework to extend the state-of-the-art (SOTA) metric BARTScore, also implemented with the 'bart-large' model. All experiments were run on a single NVIDIA v100 GPU and all metrics, apart from ROUGE, made use of the GPU compute. For the long document data set evaluation, all metrics were applied in a reference-free setting, i.e., comparing the predicted summary to the source document.

5.2. Long Document Data Set Results

To calculate the correlations between the human measure of factual consistency and automatic metrics, the fine-grained annotations were averaged to give a summary-level score. The summary-level human-annotated factuality scores were then averaged over the different annotators for each unique summary, thus giving a single summary-level human factuality score for each unique summary. These human-annotated, summary-level scores were then compared to summary-level scores predicted by each metric. Consequently, for each pair

[4] https://github.com/grrrr/krippendorff-alpha

[5] https://huggingface.co/spaces/evaluate-metric/rouge

[6] https://github.com/salesforce/factCC

[7] https://github.com/ThomasScialom/QuestEval

[8] https://github.com/neulab/BARTScore

[9] https://huggingface.co/facebook/bart-large

Randomly selected sentence from summary	Most similar sentences from source document	Your rating
for more subsequent investigations, the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti	0. in addition , single - nucleotide polymorphism (snp) analysis of fimh is a screening tool for epidemiological typing of upec (11 , 12) . therefore , the research on bacterial virulence factors can result in expansion and development of new methods for diagnosis and prevention of utis . for more subsequent investigations , the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti , referred to educational hospitals of shahrekord . 1. for more subsequent investigations , the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti , referred to educational hospitals of shahrekord . the present study was conducted for detection of the fimh virulence gene from upec isolated from both hospitalized patients and outpatients with uti , referred to educational hospitals of shahrekord , iran .	E
the fimh gene was detected in 98% of the e. coli isolates from uti.	0. for example , kaczmarek et al . (13) evaluated and detected the genes encoding virulence factors among e. coli strains with k1 antigen as well as the non - k1 e. coli strains . they found that the fimh gene existed in the whole tested e. coli k1 strains as well as in 97.0% of non - k1 strains . 1. (16) studied 18 upec isolates collected from females and found that the fimh gene was the most prevalent virulence factor and 100% of the isolates had that gene . in another study , 17) demonstrated that the fimh gene was the most frequent virulence gene and was detected in 98% of e. coli strains isolated from patient with utis .	E
we evaluated the virulence of the fimh gene in upec isolates from hospitalized patients and out - patients with upec.	0. in addition , single - nucleotide polymorphism (snp) analysis of fimh is a screening tool for epidemiological typing of upec (11 , 12) . therefore , the research on bacterial virulence factors can result in expansion and development of new methods for diagnosis and prevention of utis . for more subsequent investigations , the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti , referred to educational hospitals of shahrekord . 1. for more subsequent investigations , the fimh gene was detected in upec strains isolated from hospitalized and out - patients with uti , referred to educational hospitals of shahrekord . the present study was conducted for detection of the fimh virulence gene from upec isolated from both hospitalized patients and outpatients with uti , referred to educational hospitals of shahrekord , iran .	NE

Figure 3: Example of a PubMed summary, annotated for factual consistency by an expert human annotator to create the LongSciVerify data set. "E" indicates that a sentence is "Entailed" and "NE" indicates a sentence is "Not Entailed".

Metric	PubMed	ArXiv
ROUGE-1	0.09	0.02
ROUGE-2	0.29	0.17
ROUGE-L	0.23	0.14
BERTScore	0.24	0.27
FactCC	-0.06	-0.08
QuestEval	0.26	0.24
BARTScore	0.39	0.49
LongDocFACTScore	**0.61**	**0.61**

Table 3: Kendall's tau correlations between the human factual consistency annotations and automatic metrics for the LongSciVerify data set.

Metric	LongEval PubMed
ROUGE-1	0.15
ROUGE-2	0.26
ROUGE-L	0.22
BERTScore	0.18
FactCC	-0.14
QuestEval	0.13
BARTScore	0.22
LongDocFACTScore	**0.29**

Table 4: Kendall's tau correlations between the human factual consistency annotations and automatic metrics for the LongEval PubMed data set.

of metrics, the correlation is calculated between 45 summaries for each of the PubMed and ArXiv subsets of the LongSciVerify data set, and 40 summaries for the LongEval PubMed data set.

Table 3 gives Kendall's tau (Kendall, 1938) correlations[10] between the human measures of factuality and the automatic metrics for the LongSciVerify data sets. Table 4 gives the results of the same evaluation on the LongEval data set. Kendall's tau correlations were calculated, rather than Spearman correlations, due to being more robust for data sets with smaller sample sizes. A pairwise correlation matrix between the automatic metrics and human annotations of factuality is given in Figure 4. We restrict this plot to only the LongSciVerify PubMed data set, due to this being the data set which achieves the highest IAA score in Table 2, and is therefore likely to be the most reliable data set for evaluation. In Table 3, Table 4, and Figure 4, LongDocFACTScore, implemented by extending BARTScore, can be seen to correlate better with the human judgement of factual consistency than any other metric. Comparatively, we find that both FactCC and QuestEval show a low correlation with human judgement. BARTScore has a reasonable correlation with the human factual consistency annotations, however, since it is required to truncate the source document, we expect that it would become decreasingly correlated with human judgement as it is used to score texts of increasing length. ROUGE-2 and BERTScore perform best out of the baseline metrics evaluated, but no baseline metrics show a strong correlation with human mea-

[10] https://scipy.org

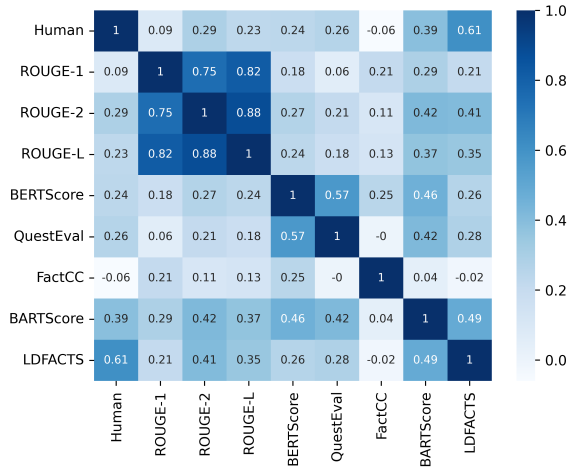

Figure 4: Pairwise Kendall's tau correlations of metrics for the LongSciEval PubMed data set. LongDocFACTScore is denoted "LDFACTS", human annotations are denoted "Human".

Metric	Time taken (s)
FactCC	24
QuestEval	160
BARTScore	**1**
LongDocFACTScore	8

Table 5: Time (s) to run each metric on 15 samples.

sures of factual consistency. Interestingly, Figure 4 shows that several automatic metrics have strong correlations with each other, suggesting that there is overlap in what they are measuring, but there is lower correlation between LongDocFACTScore and the other automatic metrics, suggesting that by providing coverage of a long document, LongDocFACTScore captures new information which the other metrics miss.

5.3. Computational Efficiency

In Table 5, we compare the average time taken, in seconds, to run each transformer-based (Vaswani et al., 2017) automatic metric designed to measure factual consistency on fifteen samples from the PubMed LongSciVerify data set. Table 5 shows that LongDocFACTScore, implemented with BARTScore, is second fastest, despite evaluating the generated summary against the entire source document, rather than a truncated version of it. In contrast, QuestEval is shown to be 20x slower, and FactCC 3x slower, than LongDocFACTScore.

5.4. Short Document Data Set Results

Although the intended use of LongDocFACTScore is to evaluate the factual consistency of abstractive summarisation for long documents, we additionally evaluate LongDocFACTScore against other automatic metrics on a variety of human annotated, short document, abstractive summarisation data sets, to validate its performance in this setting. We repeat the analysis conducted by Yuan et al., (2021) on the data sets containing human measures of factuality, and use their human annotated data and code[11], to report the Spearman correlation results for the SummEval data set's factuality measure (Fabbri et al., 2021), the accuracy scores for the Rank19 data set's factuality annotations (Falke et al., 2019), and the Pearson correlation between the automatic metrics and the human factuality annotations for the two QAGS data sets (Wang et al., 2020). We used same the measures of correlation for each data set as in the original analysis conducted by Yuan et al., (2021), rather than Kendall's tau correlations, to enable a direct comparison to their reported scores.

Table 6 gives the results of this analysis. The top section of Table 6 gives the results reported by Yuan et al., (2021), in middle section we report our results, and in the bottom section we re-report G-EVAL (Liu et al., 2023) results, a SOTA metric which uses GPT-4 to calculate factuality scores. In Table 6, it can be seen that LongDocFACTScore performs comparably in its ability to measure factual consistency to the original BARTScore model for short document summaries, indicating that the framework can be used on documents of differing lengths. G-EVAL-4 reports higher correlations results on the SummEval and QAGS-XSUM data sets but slightly lower correlations on QAGS-CNN. They do not report results for the Rank19 data set.

Although GPT-4 has a much greater token limit than standard PLMs, there is still ultimately a limit, therefore, in future work, LongDocFACTScore could be used to extend this metric, or other LLM-based metrics to very long documents, or multi-document settings. Although the token limit is longer for LLMs, non-relevant content can distract LLM-based metrics and degrade performance (Min et al., 2023), therefore we hypothesise that applying the LongDocFACTScore framework, which considers smaller sections of text at one time, could improve the performance of these models. Future work which extends LLM-based metrics should also consider in its analysis the computational cost of running an LLM-based metric in comparison to running more efficient metrics.

5.5. Parameter Study

We study effects that different parameter settings have on the LongDocFACTScore metric. We report the impact of different parameter settings on

[11] https://github.com/neulab/BARTScore

	SE Fact	R19 Acc	QAGS CNN	QAGS XSum
ROUGE-1	0.16	0.57	0.34	-0.01
ROUGE-2	0.19	0.63	0.46	0.10
ROUGE-L	0.12	0.59	0.36	0.02
MoverScore	0.16	0.71	0.41	0.05
BERTScore	0.11	0.71	0.58	0.02
FactCC	-	0.70	-	-
QAGS	-	**0.72**	0.55	0.18
BARTScore	0.31	0.68	**0.66**	0.01
LDFACTS (BARTScore)	0.36	0.68	0.65	0.04
G-EVAL-4	**0.51**	-	0.63	**0.56**

Table 6: Correlation between human measures of factuality on short document data sets, including re-reported results (Yuan et al., 2021; Liu et al., 2023). LDFACTS denotes LongDocFACTScore.

LongDocFACTScore setting	Score
BARTScore	0.440
LongDocFACTScore $K = 1$	0.605
LongDocFACTScore $K = 3$	**0.610**
LongDocFACTScore $K = 5$	0.600
LongDocFACTScore $K = 7$	0.600
LongDocFACTScore $K = 9$	0.595
LongDocFACTScore $K = 11$	0.590
LongDocFACTScore $K = I$	0.575

Table 7: The effect of varying K, the number of similar sentences considered for the LongDocFACTScore calculation, on the Kendall's tau correlation with human judgements of factuality.

the Kendall's tau correlations when evaluating the LongSciVerify data set. The PubMed and ArXiv articles are combined for this parameter study.

Table 7 shows the effect of varying K in the LongDocFACTScore framework (i.e., the maximum number of candidate similar source document sentences considered per summary sentence) on the Kendall's tau correlation with the human measures of factual consistency. The last row, $K = I$, gives the Kendall's tau correlation when all sentences in the source document are considered. The correlation of BARTScore with human annotations of factual consistency is also provided for reference. $K = 3$ is shown to be the best parameter, however, the effects of varying K are seen to be small. This is somewhat expected as the maximum score of the K text snippets is carried forward in the LongDocFACTScore metric, and it is likely that the highest scoring sentences correlate well with the most similar sentence embeddings.

Although varying K is not shown make a large difference to the performance of the metric, by selecting $K = 3$ candidate sentences, rather than cycling through all sentences in the source document (i.e., $K = I$), the score calculation in LongDocFACTScore is only calculated for approximately 1-2% of sentences from the source articles in the PubMed and ArXiv data sets. Therefore, by increasing the number of candidate similar sentences K, LongDocFACTScore becomes decreasingly efficient and, by extension, less suitable for use on long documents. To illustrate this point, in Table 8 we give the results of the repeated efficiency calculation from Table 5, where LongDocFACTScore is implemented with $K = 3$ and $K = I$. If $K = I$, there is no need to calculate sentence embeddings or perform the sentence similarity calculation, therefore we additionally report the time taken without these calculations. Table 8 shows that, for the LongSciVerify PubMed long document data set, performing the sentence similarity calculation to select the $K = 3$ most similar text snippets speeds up the metric over 15x.

LongDocFACTScore setting	Time taken (s)
$K = 3$	8
$K = I$	134
$K = I$ (no similarity calculations)	125

Table 8: Time taken (s) to run LongDocFACTScore on 15 samples, when implemented with different settings.

Method	Score
$s_k^* = s_k$	0.605
$s_k^* = s_{k-1} + s_k + s_{k+1}$	**0.610**
$s_k^* = s_{k-2} + s_k + s_{k+2}$	0.595

Table 9: The effect of varying the number of source document sentences concatenated for the LongDocFACTScore calculation on the Kendall's tau correlation with human judgements of factuality.

In Table 9, the number of candidate sentences is kept constant at $K = 3$ and the effect of concatenating the source sentence with the previous and following sentence(s) to generate a text snippet is examined on the documents from the LongSciVerify data set. Table 9 shows that although concatenating one sentence either side of a selected sentence performs best, there is little variation in the Kendall's tau correlation between the different settings.

6. Conclusion

The prevalence of LLMs and other neural methods for abstractive summarisation of long documents in real world settings is rapidly increasing, however, the abstractive methods used to generate these summaries have known issues with factual inconsistency and hallucination. In this work, we begin to address the lack of research into the suitability of automatic evaluation metrics for assessing factual consistency of long document summarisation, and make the following contributions: (i) we show that existing automatic metrics for assessing factual consistency, which have previously shown good performance on short document data sets, do not perform well in long document settings, (ii) we propose a new framework, LongDocFACTScore, which is able to consider an entire source document in its calculation, without the need to truncate it, and outperforms existing SOTA metrics in its ability to correlate with human measures of factual consistency on long document summarisation data sets, whilst still being more efficient than many SOTA automatic evaluation metrics, (iii) we work to address the lack of resources for evaluating automatic metrics in long document settings and release our LongSciVerify data set, designed for evaluating factuality metrics on the long document summarisation task. We hope that this work promotes further research into automatic metrics for evaluating abstractive summarisation of long documents. In future work, we hope to apply the LongDocFACTScore framework to extend other automatic metrics, such as newer LLM-based metrics. We also hope to incorporate our work into wider LLM evaluation frameworks.

Limitations

Firstly, we review the limitations of our human evaluation study. In our study, we recruited expert annotators, as the long document data sets are domain specific. It is difficult to recruit large numbers of expert annotators and therefore an improvement on this work would be to conduct a larger human evaluation study with more annotators evaluating more documents. We also note that two out of three annotators of the ArXiv data set have a first language which is not English, although they are both fluent in English. Furthermore, although the annotators of the ArXiv data set had all achieved a minimum of an undergraduate degree in a physical science, they did not necessarily study physics, which was the domain of most articles randomly sampled for human evaluation.

Secondly, we comment on the limitations of the LongDocFACTScore metric. One issue with this metric, and other SOTA factuality metrics, is that they favour extractive summaries. Therefore, although this metric is shown to be effective at measuring the factual consistency of long document abstractive summaries, we suggest that this metric is used in conjunction with other metrics, within a wider framework, such as FLASK (Ye et al., 2023). We also note that this evaluation only included English-language data sets.

Lastly, we discuss the computational cost of our work. We were able to monitor our GPU usage and found that for all experiments run in this period, we used approximately 1200 GPU hours. Despite our metric, LongDocFACTScore, being comparably efficient (see Table 5 and Table 8), we acknowledge that working with large neural models, as well as having environmental implications, is not economically possible for many researchers.

Ethics Statement

Throughout our research, we complied with our institution's ethical guidelines. We used open-source data and software, for which no ethical approvals were required.

In our study, we conduct a human evaluation. As detailed in Section 4, we were fortunate enough to be able to recruit colleagues, who are domain-experts in the field of the data sets. They volunteered to participate in the study without payment, so we did not need to consider the ethics of crowdworker payment.

Our work proposes a metric for assessing the factual consistency of abstractive summaries generated for long documents. This metric can be used to help researchers assess the performance of their summarisation methods, however, to minimize any harm which may be caused by deploying an abstractive summarisation model in a live setting, we suggest that any method should be thoroughly evaluated by humans in the setting it is intended to be deployed.

Acknowledgements

This work was partially supported by the project JPNP20006 from New Energy and Industrial Technology Development Organization (NEDO).

7. Bibliographical References

Reinald Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. 2022. Smart: Sentences as basic units for text evaluation.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia,

Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

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

Jennifer Bishop, Qianqian Xie, and Sophia Ananiadou. 2022. GenCompareSum: a hybrid unsupervised summarization method using salience. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 220–240, Dublin, Ireland. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

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

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.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8).

Klaus Krippendorff. 2004. Content analysis: An introduction to its methodology.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher.

2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

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.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation.

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

Shashi Narayan, 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.

OpenAI. 2023. Gpt-4 technical report.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2022. T5score: Discriminative fine-tuning of generative evaluation metrics. *arXiv preprint arXiv:2212.05726*.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *Proceedings of the American Medical Informatics Association*, 2021:605–614. Publisher Copyright: ©2021 AMIA - All rights reserved. Copyright: This record is sourced from MEDLINE/PubMed, a database of the U.S. National Library of Medicine.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey for biomedical text summarization: From pre-trained to large language models.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. 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. 2020b. Bertscore: Evaluating text generation with bert.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization.

A. LongSciVerify: Abstractive Summarisation Methods

For our human evaluation, we provide summaries generated using three different abstractive summarisation methods, which all consider text from across the entire length of a long document when generating a summary.

Text zoning is a task which aims to segment a larger body of text into different zones or sections (Teufel et al., 1999). Two of the methods we implement apply text zoning, and treat the identified sections independently, so that a PLM used to train the abstractive summarisation model is only required to process a document section at a time, rather than the entire document, to avoid truncation. We first implement DANCER (Gidiotis and Tsoumakas, 2020), a SOTA method which fine-tunes the PEGASUS PLM (Zhang et al., 2020a). DANCER splits a document into zones using keyword matching, then finds corresponding sections of the target abstract using ROUGE matching (Lin, 2004). It then uses beam search decoding to generate the summaries and combines the generated summaries of each section to form the article summary. An example output of the DANCER method can be seen in the top block of Figure 5. The second method we implement is a method we develop. It applies a similar zoning approach to DANCER, but with two notable differences. Firstly, the sections of the summary used to create the target training pairs are matched by using both keywords and assumptions about the structure of the long scientific documents used in our study (Cohan et al., 2018). Consequently, the target sentences for each section are always sentences which were consecutive to one another in the original summary, thus resulting in a summary which follows the logical structure of the document. Secondly, the summaries generated do not use beam decoding and are highly structured as each section of the summary is prefixed with the type of section it was generated from, e.g., '*results:*'. An example of a summary generated by this method can be seen in the middle block of Figure 5. The last method we implement uses an extractive-abstractive approach. We train our extractive-abstractive model using ORACLE extractive summaries as an input, optimized for a recall metric but limiting the number of sentences selected so that the total number of input tokens is less than 1024. At test time, we implement the unsupervised, extractive method GenCompareSum (Bishop et al., 2022), which does not truncate the source document and has previously shown strong performance on the PubMed and ArXiv data sets. An example of a summary generated with this method can be seen in the final block of Figure 5. We use the train/val/test splits from the original data sets and use DANCER and GenCompareSum with their default settings. For our zoning method and the extractive-abstractive method, we fine-tune the LED PLM (Beltagy et al., 2020) for three epochs with its default parameters. All experiments are run on a single NVIDIA v100 GPU.

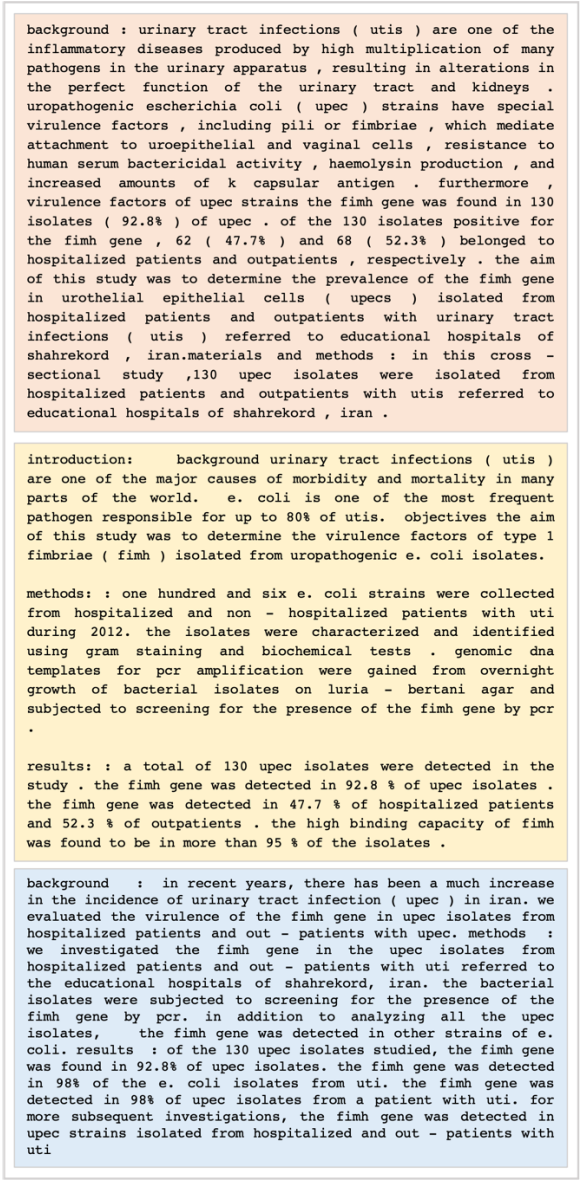

Figure 5: Automatically generated summaries included in our human evaluation study for one article sampled from the PubMed data set.