How Far are We from Robust Long Abstractive Summarization?

Huan Yee Koh[1*], **Jiaxin Ju**[4*], **He Zhang**[5], **Ming Liu**[2†], **Shirui Pan**[3†]
[1]Faculty of Information Technology, Monash University, Australia
[2]School of Information Technology, Deakin University, Australia
[3]School of Information and Communication Technology, Griffith University, Australia
[4]Independent Researcher
[5]Zhongtukexin Co. Ltd., Beijing, China
huan.koh@monash.edu, jiaxin.ju.14@gmail.com, zhanghe@kxsz.net
m.liu@deakin.edu.au, s.pan@griffith.edu.au

Abstract

Abstractive summarization has made tremendous progress in recent years. In this work, we perform fine-grained human annotations to evaluate long document abstractive summarization systems (i.e., models and metrics) with the aim of implementing them to generate reliable summaries. For long document abstractive models, we show that the constant strive for state-of-the-art ROUGE results can lead us to generate more relevant summaries but not factual ones. For long document evaluation metrics, human evaluation results show that ROUGE remains the best at evaluating the relevancy of a summary. It also reveals important limitations of factuality metrics in detecting different types of factual errors and the reasons behind the effectiveness of BARTScore. We then suggest promising directions in the endeavor of developing factual consistency metrics. Finally, we release our annotated long document dataset with the hope that it can contribute to the development of metrics across a broader range of summarization settings.

1 Introduction

Pre-trained Transformers (Devlin et al., 2019; Raffel et al., 2020) have brought tremendous progress in summarizing text in an abstract manner (Rothe et al., 2021). Unlike extractive summarization (Xiao and Carenini, 2019; Cui and Hu, 2021; Ju et al., 2021; Shi et al., 2022), abstractive summarization presents a blue-sky potential of generating summaries that are fluent and relevant to the source by intelligently paraphrasing salient contents rather than merely copying from source texts (Beltagy et al., 2020; Ju et al., 2020; Zaheer et al., 2020; Huang et al., 2021). Nevertheless, even under a short document setting, Transformer-based abstractive models often generate summaries that are repetitive (See et al., 2019; Holtzman et al., 2019), ungrammatical, and factually inconsistent with the source (Durmus et al., 2020; Kryscinski et al., 2020; Maynez et al., 2020). Furthermore, current pre-trained Transformers have an input length limit that restricts them to be directly adapted to long document summarization (Lewis et al., 2020; Zhang et al., 2020) as it would lead to a significant loss of salient information in the remaining text. These naturally bring us to a question: *How far are we from building a robust abstractive summarization system for long documents?*

A robust abstractive summarization system should at least have (i) models that can generate high-quality summaries, and (ii) evaluation metrics that can critically assess the relevance and factuality of a summary[1]. However, research on analysis and critiques of models (Wilber et al., 2021; Ladhak et al., 2022) and metrics (Gabriel et al., 2021; Pagnoni et al., 2021) mainly focus on the short-document (Kryściński et al., 2019; Fabbri et al., 2021) or long dialogue (Zhang et al., 2021). Consequently, our work aims to fill the gap by systematically analyzing abstractive models and evaluation metrics under the long document setting.

To analyze the quality of current state-of-the-art long document abstractive models, we lack a set of model-generated summaries with sufficient diversity under long document settings. To this end, we implement BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020) models under arXiv (Cohan et al., 2018) and GovReport (Huang et al., 2021) as they have been found to be the most effective pre-trained Transformer in a large-scale evaluation of summarization models (Fabbri et al., 2021). However, their 1,024 token input limit would lead to a significant loss in the information required to generate a high-quality summary.

[*] Equal contribution.
[†] Corresponding author.

[1]In machine learning parlance, robustness refers to the ability of a model to adapt to unseen distribution. Here, robustness refers to the effectiveness of a model to adapt from short to long documents to generate relevant and factual summaries.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2682 - 2698
December 7-11, 2022

Hence, by closely following prior works in extending the pre-trained models using sparse attention (Beltagy et al., 2020; Zaheer et al., 2020; Huang et al., 2021) and reduce-then-summarize mechanism (Pilault et al., 2020; Zhang et al., 2022), we implement different variants of Longformer-based BART and PEGASUS to obtain a diverse set of summaries. We then perform fine-grained human analysis on the model outputs by three human annotators to qualitatively assess whether long document abstractive models can generate relevant and factually consistent summaries.

Effective evaluation metrics are also paramount as they can critically assess the model performance before releasing it to target users. We adapt recently proposed metrics (Durmus et al., 2020; Kryscinski et al., 2020; Nan et al., 2021; Yuan et al., 2021; Laban et al., 2022) to long document settings and thoroughly analyze their strength and weaknesses to measure the relevance and factual consistency on our annotated dataset. To our best knowledge, we are the first to assess abstractive models and evaluation metrics under the long document setting.

Our contributions are as follows: (1) We analyze pre-trained Transformer summarizers to encourage a rethinking of architectural designs under long document settings. (2) We release human-annotated long document abstractive model outputs to further research in human-correlated evaluation metrics across a broader setting. (3) We investigate summarization metrics using our annotated long document datasets to expose the limitation of metrics and provide promising directions for the future development of evaluation metrics.

2 Related Work

2.1 Long Abstractive Models

To implement pre-trained Transformers (Devlin et al., 2019; Raffel et al., 2020) for long document summarization tasks, they have to be adapted with *long document mechanisms* to improve models' efficiency and extend their input limit (Koh et al., 2022). In this work, we focus on analyzing abstractive models after incorporating the two following long document mechanisms:

Sparse Attention It aims to reduce the quadratic complexity of Transformers into sub-quadratic complexity (Child et al., 2019; Kitaev et al., 2019; Choromanski et al., 2020) while exploiting the benefits of pre-training (Beltagy et al., 2020; Zaheer et al., 2020; Huang et al., 2021; Guo et al., 2022; Pietruszka et al., 2022). The gain in efficiencies allows Transformer to be fine-tuned on downstream summarization tasks with a substantially longer input text. Despite a plethora of proposals on sparse attention, Xiong et al. (2022) recently showed that simple local attention remains competitive.

Reduce-then-Summarize This approach aims to reduce the source text into a shorter subset, allowing it to fit within the input token limit of a Transformer. The source text can be reduced into a more condensed text through extraction of salient sentences (Pilault et al., 2020; Zhao et al., 2020; Bajaj et al., 2021) or generation of shorter texts from segments of the source (Gidiotis and Tsoumakas, 2020; Zhang et al., 2022). These models often train Transformer-based summarizers using reduced source texts which greedily maximize ROUGE scores and utilize separate retrievers during the testing stage to avoid "cheating" (Pilault et al., 2020; Manakul and Gales, 2021; Mao et al., 2022). Importantly, the retriever will also be trained to maximize ROUGE to avoid a significant disconnect between the training and testing stage.

2.2 Evaluation Metrics

Given the limitations of the ROUGE metric (Chaganty et al., 2018; Kryściński et al., 2019), new metrics are proposed to better measure two fundamental qualities of summary: relevance and factual consistency. Relevance metrics such as ROUGE variants (Ng and Abrecht, 2015; Ganesan, 2018; ShafieiBavani et al., 2018) and BERTScore (Zhang et al., 2019) measure whether a summary contains the main ideas of the source. A factual consistency metric assesses whether a summary is factually consistent with the source (Goyal and Durrett, 2020; Wang et al., 2020). Due to the high rate of factual errors in the summaries generated by short-document models (Cao et al., 2018; Maynez et al., 2020), there have been substantial efforts in developing effective metrics which can measure the factuality of a summary (Honovich et al., 2021; Xie et al., 2021; Ribeiro et al., 2022).

3 Generation of Model Summary

To investigate the robustness of long document abstractive systems, we need a set of model-generated summaries that can roughly represent the state of current research progress. In this section, we describe our methodology to obtain such samples.

3.1 Model Variants

Pretraining Task We implement BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020). Both models have a 1,024 input token limit with extra text tokens to be truncated. We extend the input limit of BART and PEGASUS using the sparse attention and reduce-then-summarize mechanism.

Sparse Attention We extend the input limit of the pre-trained Transformer using Longformer's adaptation to have a maximum input of 1K, 4K, and 8K tokens (Beltagy et al., 2020). Xiong et al. (2022) recently showed that local-window attentions (i.e., only attending to neighborhood tokens) are sufficient and competitive against other variants. The Longformer sparse attention adaptation thus gives us a reasonable baseline representation for current long document abstractive summarizers.

Reduce-then-Summarize To explore the effectiveness of the reduce-then-summarize approach, we implement an oracle retriever by greedily extracting salient sentences that maximize ROUGE-2 up to the input limit of Transformer during the training and inference stage. Although using reference summaries to extract the salient sentences at the testing stage is considered cheating, contemporary approaches are trained to retrieve oracle summaries and are thus trained to become an oracle retriever (Manakul and Gales, 2021; Mao et al., 2022). Using an oracle retriever allows us to analyze whether the reduce-then-summarize approach will generate desirable summaries given that the retriever is *perfectly* trained with its upper bound performance. This allows us to analyze whether the summary generated from a ROUGE-maximizing model with a reduce-then-summarize mechanism will be desirable for target users. We implement models with 1K, 4K, and 8K tokens of the reduced subset.

3.2 Long Document Dataset

We implement the model configurations above on the ArXiv (Cohan et al., 2018) and GovReport (Huang et al., 2021) because they cover a wide range of topics in the scientific and general domains respectively. Both have an average source length of greater than 6,000 tokens, sufficiently long to challenge pre-trained Transformers. Besides, arXiv requires models to paraphrase more as compared to GovReport. Both datasets are chosen after analyzing the characteristics of datasets across 10 benchmark datasets with details in Appendix A.5.

Kryściński et al. (2019) has shown that 60% of most important sentences lie within the leading one-third of the CNN-DM articles (Nallapati et al., 2016). However, the linguistic styling and structure of a short document would often differ significantly from a long document. To investigate how much information a model would lose when processing only the leading text, we plot the distribution of salient content of arXiv and GovReport. This is done by performing human annotation on 10% randomly sampled document-summary pairs from arXiv (700) and GovReport (100) test set. For each sentence in the reference summaries, we trace back the leading source paragraph position in the original document that contains the idea required to generate a sentence.

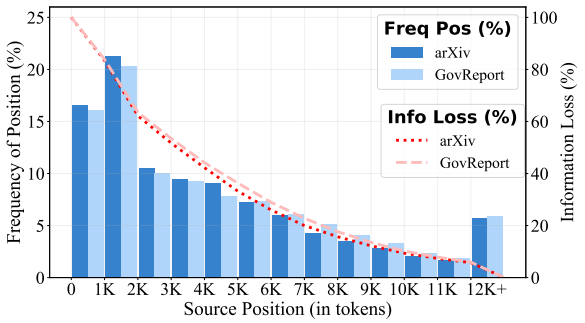

Figure 1: Distribution of salient content against the document length according to human annotators (left) ; Information loss of Transformer-based abstractive summarizers based on different input limits (right).

Distribution plot in Figure 1 shows the source position frequency in terms of the total percentage of the occurrence. The line plots illustrate the total information loss given an input limit. This reflects the information loss of a model when it only takes the leading source tokens. The line plot suggests that an input limit of 1K, 4K, and 8K tokens would equate to roughly 80%, 40%, and 20% average information loss respectively on both datasets.

Importantly, we see more salient information to be distributed from 1K to 2K tokens than 0 to 1K tokens, suggesting that the strategy of vanilla BART and PEGASUS to process the leading 1K input limit is sub-optimal. We hope that the result here would also provide directions for future architectural designs to identify salient contents.

3.3 Training Details

Given two pre-training tasks with three input limit settings for Longformer-based Sparse Attention and Reduce-then-Summarize settings, this gives us

	Summarization Model	Longformer	InputType	InputLen	InfoLoss	arXiv R-1	arXiv R-2	arXiv R-L	GovReport R-1	GovReport R-2	GovReport R-L
Sparse Attention	BART (LEAD 1K)	x	**LEAD**	**1K**	80%	43.84	16.55	39.86	56.55	26.70	54.46
	BART (LEAD 4K)	✓	**LEAD**	**4K**	40%	45.72	18.48	41.82	57.45	28.14	55.40
	BART (LEAD 8K)	✓	**LEAD**	**8K**	20%	46.60	19.05	42.21	58.35	28.78	**56.35**
	PEGASUS (LEAD 1K)	x	**LEAD**	**1K**	80%	44.17	17.16	40.18	57.19	27.87	55.17
	PEGASUS (LEAD 4K)	✓	**LEAD**	**4K**	40%	46.02	18.33	42.28	58.35	28.78	56.35
	PEGASUS (LEAD 8K)	✓	**LEAD**	**8K**	20%	**46.87**	**19.73**	**42.36**	**58.59**	**29.02**	56.29
Reduce-then-Summ	BART (ORACLE 1K)	x	**ORACLE**	**1K**	-	50.43	**24.16**	44.93	63.07	36.64	60.09
	BART (ORACLE 4K)	✓	**ORACLE**	**4K**	-	49.75	23.05	44.41	60.21	31.34	57.13
	BART (ORACLE 8K)	✓	**ORACLE**	**8K**	-	49.13	21.52	44.72	59.06	29.66	56.37
	PEGASUS (ORACLE 1K)	x	**ORACLE**	**1K**	-	**50.50**	23.59	**45.03**	**63.47**	**37.27**	**60.52**
	PEGASUS (ORACLE 4K)	✓	**ORACLE**	**4K**	-	46.21	20.32	42.23	60.86	33.68	57.88
	PEGASUS (ORACLE 8K)	✓	**ORACLE**	**8K**	-	49.06	20.60	43.55	58.77	31.53	56.51
SOTA	TDT (Pang et al., 2022)	**#N/A**	**LEAD**	**16K**	-	**50.95**	21.93	**45.61**	-	-	-
	DYLE (Mao et al., 2022)	**#N/A**	**DYNAMIC**	**#N/A**	-	46.41	17.95	41.54	61.01	28.83	57.82

Table 1: ROUGE score validation of implemented pre-trained BART and PEGASUS. SOTA stands for current state-of-the-art on arXiv, TDT (Pang et al., 2022), and GovReport, DYLE (Mao et al., 2022). **Red** represents best dataset result and **Bold** represents best result under the sparse attention or reduce-then-summarize setting.

12 model configurations per dataset. For 1K token configurations, we use BART-large and PEGASUS-large. For 4K and 8K token configurations, we follow Longformer's implementation in extending the position embedding to 4K and 8K tokens by repeatedly copying position embeddings of BART and PEGASUS. To ensure comparability, all 24 models have a fixed output length of 512 tokens and are fine-tuned independently on RTX 3090 GPU with 24 GiB of GPU memory. We follow original authors in train/validation/test split of ArXiv (Cohan et al., 2018) and GovReport (Huang et al., 2021). Implementation details in Appendix A.3.

3.4 ROUGE Validation

Table 1 shows that sparse attention models achieve competitive but lower ROUGE than state-of-the-art models, arXiv-TDT (Pang et al., 2022) and GovReport-DYLE (Mao et al., 2022). Extending the vanilla BART and PEGASUS using Longformer also provides a performance boost as the information loss is reduced exponentially when the input limit increased from 1K to 4K and 8K. The reduce-then-summarize models achieve ROUGE that either match or exceed arXiv's and GovReport's state-of-the-art. As increasing the input length would place more burden on reduce-then-summarize models to identify tokens that maximize ROUGE over long sequences, we see a slight decrease in ROUGE as the length is increased.

The above results indicate that the implemented Longformer-based sparse attention models can reasonably reflect the current long abstractive summarization baselines, while the reduce-then-summarize models can roughly represent the summary outputs of state-of-the-arts under arXiv and GovReport. In the next two sections, we will investigate whether the advancement in summarization research has brought us far enough to build a robust summarization system (i.e., model and metric) based on the summaries generated from all of the 24 implemented summarizers in this section. For consistency, we will refer to Longformer-based sparse attention BART and PEGASUS as BART/PEGASUS (LEAD #K) as it only takes the leading input token, whereas, reduce-then-summarize models will be referred to as BART/PEGASUS (ORACLE #K). The # symbol represents the token input length limit of the Transformer-based summarizer.

4 Human Evaluation of Models

To assess the overall quality of summaries, we randomly sampled 204 model-generated summaries from each dataset to be evaluated by three annotators based on the relevance and factual consistency aspect. To ensure comparability between model variants, we randomly sampled document IDs from the test set and extracted all 12 corresponding model summaries to annotate. As each summary ranged from 5 to 15 sentences, we annotated 4,190 sentences, matching a large-scale human evaluation by Pagnoni et al. (2021) of 2,250 short-document articles.

4.1 Annotation Procedures

Relevance Relevance measures whether a summary contains the main ideas of the source. As the author is arguably the best person to summarize the source, we assign relevance scoring based on the percentage of the reference summary's main ideas contained in the generated summary. The relevance score of each summary is the average of three annotation samples.

Factual Consistency Factual consistency measures whether a candidate summary is factually consistent with the source. Following Pagnoni et al. (2021), we classify each summary sentence's factuality based on seven types of errors: i) *PredE* - predicate in summary inconsistent with source, ii) *EntityE* - primary arguments or its attributes are wrong, iii) *CircE* - predicate's circumstantial information is wrong, iv) *CorefE* - co-reference error, v) *LinkE* - multiple sentences linked incorrectly, vi) *OutE* - out of article error and vii) *GramE* - unreadable sentence(s) due to grammatical errors. Similarly, the factual consistency of a summary is the percentage of factually consistent sentences and the final score is the average of three samples.

Inter-Annotator Agreement Following Fabbri et al. (2021), the inter-annotator interval kappa of relevance score between the three annotators is 0.5874, computed based on Krippendorff's alpha coefficient (Krippendorff, 2011) where each score is assigned to a multiple of quarter intervals. To calculate inter-annotator agreement of factual consistency, we follow Durmus et al. (2020); Pagnoni et al. (2021) in using Fleiss Kappa, κ, and the percentage, p, of annotators that agree with the majority class. With a total of 4190 sentences, we observe $\kappa = 0.52$ and $p = 84\%$, slightly lower but comparable to Pagnoni et al. (2021)'s result ($\kappa = 0.58$ and $p = 91\%$).

4.2 Long Abstractive Model Analysis

Relevance Benefiting from processing the oracle inputs, Figure 2 shows BART/PEGASUS (ORACLE #K) to achieve a higher relevance score than BART/PEGASUS (LEAD #K). On average, PEGASUS also performs better than BART. Looking at the models with the same pre-training tasks, we observe that BART (ORACLE #K) did not significantly outperform BART (LEAD #K) on arXiv. On the other hand, PEGASUS (ORACLE #K) shows a significant improvement over PEGASUS (LEAD #K) under both the arXiv and GovReport dataset. We hypothesize that when models take the oracle inputs, the text becomes incoherent and the immediate connection between sentences is less obvious, causing it harder for BART models to understand the contextual dependencies between the tokens. In contrast, PEGASUS's Gap-Sentence Generation pre-training may help models in reasoning the contextual dependencies of an incoherent text.

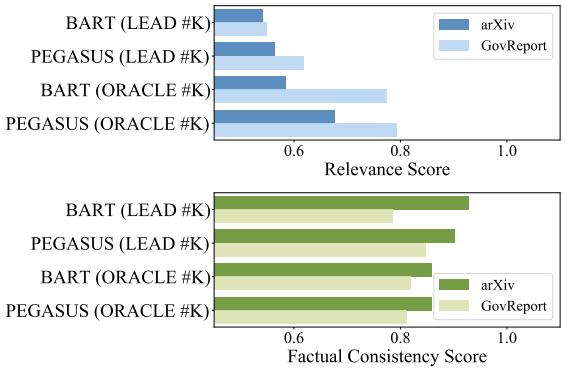

Figure 2: *Average* human relevance (top) and factual consistency (bottom) scores for BART and PEGASUS models with 1K, 4K and 8K input limit.

Factual Consistency On average, we also observe PEGASUS makes fewer factual errors as compared to BART across most settings. Unlike the relevance aspect, the BART/PEGASUS (ORACLE #K) setting often achieves lower factual consistency results as compared to BART/PEGASUS (LEAD #K). This indicates that while models can more easily capture relevant text, incoherent texts may cause them to make more factual errors. As BART/PEGASUS (ORACLE #K) utilize an oracle retriever during testing that is not allowed under normal settings, similar issues could potentially be exacerbated when a model-based retriever (Pilault et al., 2020; Manakul and Gales, 2021; Mao et al., 2022) is used to extract salient sentences from the source. Finally, this also indicates that maximizing ROUGE itself leads us to models with more relevant summaries but may not be necessarily factual.

Summary Quality v.s. Input Limit Other than high-level analysis of the different pre-training and mechanism results, we investigate the relationship of the adjustment in input limit of different Transformer variants against the human-annotated relevance and factual consistency scores. Table 2 shows that the relevance score increases when the input limit of the BART/PEGASUS (LEAD #K)

Model Configuration	arXiv		GovReport	
	REL	FACT	REL	FACT
BART (LEAD #K)	$+2.78^{**}$	$+1.98^{*}$	$+3.03^{**}$	$+1.03$
PEGASUS (LEAD #K)	$+0.31$	$+0.81$	$+1.95^{**}$	$+0.67$
BART (ORACLE #K)	-0.52	-0.29	-0.61	$+0.03$
PEGASUS (ORACLE #K)	$+0.13$	-0.14	$+0.21$	$+0.12$

Table 2: Coefficient of simple linear regression of Relevance and Factual Consistency against Input Limit. Values are in percentage point per 1K input limit. * represents $p < 0.05$ and ** represents $p < 0.001$.

models is extended but does not show meaningful differences when the oracle input length of the BART/PEGASUS (ORACLE #K) models is adjusted. Since longer oracle input length increases the difficulty of identifying salient content for a BART/PEGASUS (ORACLE #K) model and the increase in difficulties did not lead to a drop in summarization performance, this suggests that both pre-trained Transformers are capable of reasoning through long-range texts. This also indicates that the gain in relevance score mostly comes from the reduction in information loss caused by the input limit of Transformer-based summarizers.

While we see an improvement in factual consistency scores when vanilla pre-trained Transformers increase their input limits using Longformer, only BART (LEAD #K) under arXiv shows a statistically significant result. The BART/PEGASUS (ORACLE #K) models do not show conclusive results as to which configurations will generate summaries that are most factually consistent.

4.3 Fine-grained Analysis of Factual Errors

Under real-world scenarios, a model will not be evaluated based on the percentage of factual sentences and is only considered robust if it generates summaries that are almost entirely error-free. However, the models generate factually *inconsistent* summaries, on average, 35% and 81% of the time under arXiv and GovReport respectively. The least errors are made by PEGASUS (LEAD 8K) in arXiv (21%) and PEGASUS (ORACLE 1K) in GovReport (60%). Given the unacceptably high amount of factual errors, it is fair to conclude that the models are not sufficiently robust. Thus, it is more important that we analyze the type of errors they made and how we can improve their performance in the factuality aspect. To this end, we investigate the proportion of summaries with different types of factual error instances in Figure 3.

As arXiv articles are pre-processed when the dataset was introduced by Cohan et al. (2018) while GovReport articles closely resemble the original documents (Huang et al., 2021), the task is made less challenging under arXiv, and mistakes related to CorefE, EntE and CircE are greatly reduced. Still, models under the arXiv setting generate higher LinkE errors as they are required to paraphrase the source text more. We also see BART (ORACLE #K) and PEGASUS (ORACLE #K) to make more LinkE errors as the oracle input text is less coherent as compared to the leading input text. We again observe PEGASUS makes fewer errors as compared to BART. The better performance of PEGASUS mostly comes from making fewer CorefE, EntE, GramE and PredE errors.

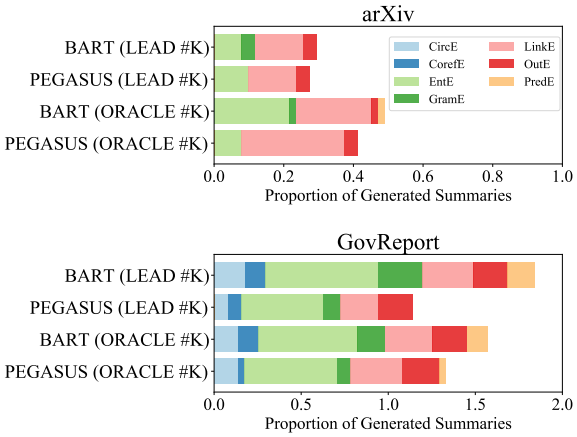

Figure 3: *Average* Proportion of Factual Error Type for all generated summaries of BART and PEGASUS models with 1K, 4K, and 8K input limits. As a long document summary have multiple sentences and can have multiple error types, the total proportion may exceed 1.

We conclude this section by noting that while ROUGE scores show minor differences between BART and PEGASUS, human evaluation relevance and factual consistency scores reveal that PEGASUS is considerably better than BART. This conflicts with the findings of Rothe et al. (2021) that PEGASUS task-specific pre-training did not bring improvement in summarization performances, emphasizing the need of evaluating summaries based on the quality judged by a summary user rather than solely relying on the ROUGE metric.

4.4 Dataset for Metrics Evaluation

We release the human-annotated summaries to encourage the future exploration of long document models and metrics[2]. In the next section, we utilize the dataset to assess long document metrics.

[2]https://github.com/huankoh/How-Far-are-We-from-Robust-Long-Abstractive-Summarization

Metrics	Relevance								Factual Consistency							
	arXiv				GovReport				arXiv				GovReport			
	Pearson		Spearman		Pearson		Spearman		Pearson		Spearman		Pearson		Spearman	
	ρ	p-val	r	p-val	ρ	p-val	r	p-val	ρ	p-val	r	p-val	ρ	p-val	r	p-val
BLEU	0.21	0.00	0.21	0.00	0.37	0.00	0.35	0.00	-0.05	0.48	-0.05	0.45	-0.12	0.09	-0.14	0.11
METEOR	0.26	0.00	0.22	0.00	0.40	0.00	0.38	0.00	0.08	0.24	0.09	0.18	-0.09	0.14	-0.13	0.12
ROUGE-1	**0.29**	0.00	**0.25**	0.00	**0.53**	0.00	**0.52**	0.00	-0.08	0.26	-0.13	0.16	-0.12	0.09	-0.11	0.12
ROUGE-2	0.14	0.03	0.16	0.02	0.43	0.00	0.44	0.00	-0.12	0.09	-0.13	0.10	-0.08	0.32	-0.11	0.10
ROUGE-L	0.12	0.07	0.17	0.06	0.38	0.00	0.39	0.00	-0.16	0.09	-0.15	0.07	-0.08	0.21	-0.11	0.11
BERTS	0.22	0.00	0.18	0.00	0.38	0.00	0.38	0.00	-0.09	0.12	-0.10	0.10	0.00	0.95	-0.04	0.57
BARTS-ZS	0.06	0.44	0.12	0.09	0.19	0.00	0.25	0.00	0.25	0.00	0.24	0.03	0.17	0.06	0.06	0.02
BARTS-FT	0.00	0.98	0.03	0.64	0.18	0.00	0.24	0.00	**0.32**	0.00	**0.36**	0.02	**0.51**	0.00	**0.48**	0.00
OpenIE	0.21	0.00	0.23	0.00	0.03	0.60	0.01	0.88	0.20	0.00	0.15	0.03	0.33	0.00	0.34	0.00
MNLI-TE	0.03	0.72	0.03	0.69	0.08	0.27	0.05	0.45	-0.08	0.19	-0.04	0.56	-0.14	0.18	-0.13	0.20
FactCC	0.13	0.07	0.13	0.07	0.05	0.52	0.04	0.55	0.22	0.00	0.19	0.00	0.28	0.00	0.27	0.00
FEQA	0.03	0.66	0.01	0.83	0.09	0.20	0.10	0.14	0.06	0.36	0.05	0.45	-0.08	0.24	0.00	0.46
QUAL	-0.06	0.38	-0.07	0.34	0.30	0.00	0.34	0.00	0.12	0.07	0.16	0.02	0.12	0.08	0.10	0.11
SummaC	0.09	0.22	0.08	0.24	0.05	0.49	0.04	0.57	**0.32**	0.00	0.32	0.00	0.39	0.00	0.38	0.00

Table 3: Statistical Relationship between human judgment (relevance and factual consistency) and metric scores based on Pearson correlation, ρ, and Spearman rank correlation, r, coefficients and their p-values. Upper and lower part show results for general metric and factual consistency metric respectively.

5 Human Evaluation of Metrics

With high factual inconsistency rates, long abstractive summarizers remained unready for real-world implementation. It is thus paramount to ensure that the performances of future proposed models can be evaluated by metrics that are well correlated with user judgment. However, previous works on evaluation metrics have mainly focused on short document summarization research settings due to (i) the lack of human-annotated long document model-generated summaries and (ii) the reliance of metrics on pre-trained language models that are fine-tuned on short document datasets (Maynez et al., 2020; Durmus et al., 2020; Wang et al., 2020). Relying on our annotated dataset, we adapt evaluation metrics proposed in prior works to the long document settings and correlate their metric scores with average human relevance and factual consistency scores.

General Metric General metrics attempt to capture the overall summary quality including relevance and factual consistency. Assessed general metrics are: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021). We implement zero-shot (BARTS-ZS) and fine-tuned (BARTS-FT) BARTScore. BART-ZS uses the original BART model while BART-FT is fine-tuned on the arXiv and GovReport datasets. Both are extended to 8K tokens using Longformer.

Factual Consistency Factual consistency metrics we assess are: OpenIE (Goodrich et al., 2019) that extracts semantic triples from source and summary, then compute scores through embedding matching (Reimers and Gurevych, 2019). FactCC (Kryscinski et al., 2020) adopts a weakly-supervised model approach. FEQA (Durmus et al., 2020) and QUAL (Nan et al., 2021) evaluate factuality using a question-generation and answering (QGA) approach. TE-MNLI (Maynez et al., 2020) and SummaC (Laban et al., 2022) are text entailment approach, TE-MNLI evaluates probability of entailment at the document-level while SummaC at the sentence-level. For metrics with short input limits, we extend the input limit of FactCC using Longformer and use the oracle summaries as a substitute for the source for FEQA, QUAL and TE-MNLI. Implementation details in Appendix A.4.

5.1 Overall Result

Relevance Contrary to past research under short-document setting (Kryściński et al., 2019; Bhandari et al., 2020; Akter et al., 2022), Table 3 shows that ROUGE scores still correlate best with the human judgment of relevance score in our settings. This provides comfort for future research to rely on the ROUGE metric for benchmarking long document abstractive models in generating relevant summaries. We hypothesize that the effectiveness of ROUGE metric is due to the linguistic styling of long document datasets that are often written in formal languages. We caution that similar results may not be achieved by ROUGE metric when the dataset and model-generated summaries are sufficiently abstractive.

Factual Consistency The metrics that achieve the best overall correlation with the human factual consistency scores are fine-tuned BARTScore, followed by SummaC, FactCC, and OpenIE. Interestingly, zero-shot BARTScore also achieves third and fifth-best results on arXiv and GovReport respectively. Consistent with Pagnoni et al. (2021), QGA approaches do not seem to achieve statistically significant results, except for QUAL under GovReport. From the perspective of efficiencies, BARTScore and FactCC require approximately 4 days of fine-tuning per dataset on an RTX 3090 GPU while zero-shot SummaC and OpenIE can be implemented immediately without dataset-specific training. On balance, SummaC and BARTS-FT seem to stand out from the rest as the most effective zero-shot and fine-tuned metric respectively. Nevertheless, it is more important to thoroughly investigate why and when the metrics will identify factual inconsistencies in model outputs.

5.2 Identification of Factual Error Types

Overall correlation with human factual consistency score does not reveal the limitations of a metric in identifying different types of factual errors (Goyal and Durrett, 2021; Pagnoni et al., 2021). Hence, we plot the contribution of each error type to the overall correlation in Figure 4. It shows the change in correlation when the error type is excluded from the calculation. As compared to Table 3, a higher positive bar value shows that the error type contributed more to the metric performances, causing a decrease in overall correlation.

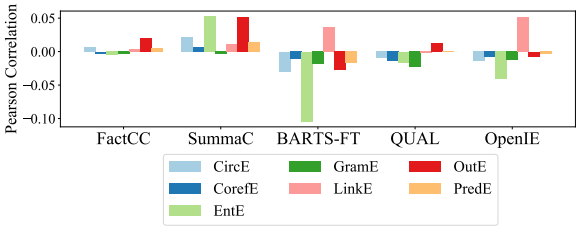

Figure 4: Change in Pearson correlation when error types are omitted. Higher value indicates a greater influence of the error type on overall correlation result.

Figure 4 shows that OpenIE and BARTScore are not able to identify entity errors (EntE) well. We hypothesize that this is because OpenIE relies on the soft-embedding similarity while BARTScore finds reasonableness in generating closely related entities in the source document. Nevertheless, BARTScore and OpenIE show better ability at identifying sentence linkage (LinkE) errors as BARTScore takes the full context of the entire generated summary into account while OpenIE assesses the relationship between semantic triples. FactCC, SummaC and QUAL which only relied on sentence- or question-level granularity did not see a high correlation with LinkE as they do not take the overall contexts of the generated summaries. SummaC shows strong correlations with entity (EntE) and out-of-article (OutE) errors. As different metrics can better identify different factual error types, combining the advantages of various metrics to address their limitations may be worthwhile. For a simple illustration, by taking the average normalized metric scores of BARTS-FT and SummaC, we are able to increase Table 3's best Pearson correlation result of arXiv from 32% to 38% and GovReport from 51% to 59%, representing an absolute percentage point increase of 6% and 8% respectively.

5.3 On the Effectiveness of BARTScore

Given the superiority of BARTScore as a factuality metric, we further analyze it in detail. BARTScore relies on a BART's average log-likelihood of generating the evaluated summary conditional on the source document: $\frac{1}{m}\sum_{t=1}^{m} \log p(\mathbf{y_t}|\mathbf{y_{<t}}, \mathbf{d})$ where $\mathbf{y_t}$ represent generated tokens in the summary at generation step t while $\mathbf{d}$ represents source (Yuan et al., 2021). Under the fine-tuned variant, BARTScore is fine-tuned as a summarization model. Thus, a lower BARTScore indicates that the BART model shows a lower likelihood of generating the evaluated texts. This suggests that summarization models are "aware" of potentially making factual errors in the form of lower generation probability. Similar to our findings, Xu et al. (2020) has found that lower generation probability (and higher entropy value) leads to greater novelty in the tokens generated but a higher chance of factual inconsistencies under short-document settings. Consequently, solving the factuality aspects of abstractive models and metrics from this perspective may be a fruitful direction to explore.

In addition, we fine-tuned BARTScore on different datasets and compute its correlation with human factual consistency scores in Table 4. BART shows a better correlation when metrics are fine-tuned on in-domain datasets. In particular, we find the best results are achieved for arXiv when BART is fine-tuned on arXiv or PubMed and for GovReport when BART is fine-tuned on GovReport.

To validate this hypothesis, we further imple-

Variants	arXiv		GovReport	
	Pearson	Spearman	Pearson	Spearman
Zero-Shot	0.25	0.24	0.23	0.17
arXiv	0.32	0.36	0.19	0.18
GovReport	0.31	0.21	**0.51**	**0.48**
PubMed	**0.34**	**0.36**	0.38	0.38
BookSum	0.31	0.31	0.36	0.31

Table 4: Human Factual Consistency correlation with BARTScore variants fine-tuned on different datasets. All results are statistically significant, where $p < 0.05$.

ment FEQA with Sci-BERT (Beltagy et al., 2019) fine-tuned on SQuaD (Rajpurkar et al., 2016, 2018) and QUAC (Choi et al., 2018) and we obtain statistically significant Pearson correlation ($\rho = +0.22$) on arXiv, a four-fold increase as compared to the original variant. This finding strongly emphasizes the importance of fine-tuning metrics on in-domain datasets. Future work on metrics could thus benefit from incorporating fine-tuning strategies (Kryscinski et al., 2020; Laban et al., 2022) rather than relying merely on publicly available models (Maynez et al., 2020; Durmus et al., 2020). Importantly, the fine-tuning strategy should be efficient and generalizable to other domains to ensure that it is not limited to short news articles.

6 Conclusion

In this work, we perform human evaluations of model-generated summaries to critically analyze the relevance and factual consistency aspect of models and metrics under long document settings.

For models, we highlight that the constant strive for higher ROUGE scores leads us to long document models with more relevant summaries but not necessarily factual ones. We also show that PEGASUS pre-training allows long document Transformer to make fewer factual errors and can comprehend incoherent text better, suggesting that PEGASUS can be more beneficial than BART for reduce-then-summarize architectures that are common for long document summarizers. For metrics, we observe that ROUGE remains superior at assessing the relevance of summaries, while a fine-tuned BARTScore can be most effective in evaluating the factuality of long document summaries.

We also release the annotated dataset to encourage analysis of summarization systems across a broader range of settings. We hope that this work can provide practical insights into future research to develop long document summarization systems that can be relied upon in our daily lives.

Limitations

Our findings and conclusions relied on human annotation efforts by three annotators. To balance the quality and quantity of annotation, three annotators evaluated the same 408 summary-document pairs across two datasets. While having three annotations per summary-document pair reduces the variability and enhances the final quality of annotation, increasing the size and diversity of our annotated dataset would further enhance the statistical significance of our findings.

Prior works on summarization metrics have assessed their performances on short summary-document pairs and often relied on pre-trained models with token limits that cannot be easily extended. While we have taken reasonable steps in adapting their methods to long document settings, it is plausible that better adaptation approaches can be discovered.

Finally, our experiments are conducted on the arXiv and GovReport benchmark datasets. The documents in both datasets are written in formal language. While formal language is common across long document benchmark datasets, this may result in domain bias. Our experimental processes and findings may also be limited to the English language. This is especially the case for our human-annotation process as we relied on English grammatical rules to determine the qualitative aspects of model-generated summaries. Thus, our processes and findings are likely not applicable to long documents that are not written in English. Nevertheless, we hope that our work can indirectly inspire or be extended to the research in multilingual long document summarization.

References

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022. Revisiting automatic evaluation of extractive summarization task: Can we do better than rouge? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1547–1560.

Ahsaas Bajaj, Pavitra Dangati, Kalpesh Krishna, Pradhiksha Ashok Kumar, Rheeya Uppaal, Bradford Windsor, Eliot Brenner, Dominic Dotterrer, Rajarshi Das, and Andrew McCallum. 2021. Long document summarization in a low resource setting using pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 71–80.

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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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 3615–3620.

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

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. In *International Conference on Learning Representations*.

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.

Peng Cui and Le Hu. 2021. Sliding selector network with dynamic memory for extractive summarization of long documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5881–5891.

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.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

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.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

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.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 166–175.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462.

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.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2:: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436.

Jiaxin Ju, Ming Liu, Longxiang Gao, and Shirui Pan. 2020. Monash-summ@ longsumm 20 scisummpip: An unsupervised scientific paper summarization pipeline. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 318–327.

Jiaxin Ju, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan. 2021. Leveraging information bottleneck for scientific document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4091–4098.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of reddit posts with multi-level memory networks. 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 2519–2531.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2019. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

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.*

Anastassia Kornilova and Vlad Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. *EMNLP-IJCNLP 2019*, page 48.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *ArXiv*, abs/1810.09305.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Wojciech Kryściński, 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.

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.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Potsawee Manakul and Mark Gales. 2021. Long-span summarization via local attention and content selection. 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 6026–6041.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Awadallah, and Dragomir Radev. 2022. DYLE: Dynamic latent extraction for abstractive long-input summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1687–1698.

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.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, 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.

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.

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.

Jun Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930.

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.

Bo Pang, Erik Nijkamp, Wojciech Kryściński, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2022. Long document summarization with top-down and bottom-up inference. *arXiv preprint arXiv:2203.07586*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Michał Pietruszka, Łukasz Borchmann, and Łukasz Garncarek. 2022. Sparsifying transformer models with trainable representation pooling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8616–8633.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Christopher Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

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.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253.

Sascha Rothe, Joshua Maynez, and Shashi Narayan. 2021. A thorough evaluation of task-specific pretraining for summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 140–145.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. A graph-theoretic summary evaluation for rouge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 762–767.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Kaile Shi, Xiaoyan Cai, Libin Yang, Jintao Zhao, and Shirui Pan. 2022. Starsum: A star architecture based model for extractive summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3436–3449.

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.

Matt Wilber, William Timkey, and Marten van Schijndel. 2021. To point or not to point: Understanding how abstractive summarizers paraphrase text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3362–3376.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. 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 3011–3021.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110.

Wenhan Xiong, Barlas Oguz, Anchit Gupta, Xilun Chen, Diana Liskovich, Omer Levy, Scott Yih, and Yashar Mehdad. 2022. Simple local attentions remain competitive for long-context tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1975–1986.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297.

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. In *International Conference on Learning Representations*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summ^n: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Findings.*

Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020. Seal: Segment-wise extractive-abstractive long-form text summarization. *arXiv preprint arXiv:2006.10213*.

A Appendices

A.1 Broader Impacts

Abstractive models implemented are in general neural conditional generation models that have a wide range of capabilities due to their ability to carry out arbitrary language generation tasks. This may have a negative society such as generating texts that are biased towards certain minorities or unfairly discriminate against a certain group. This risk may, for example, arise from the human-annotated model dataset that we aim to release along with this work. Nevertheless, we have taken sufficient care to ensure that the potential risks of broad negative impacts are minimized. Based on our annotation, we believe that the risks of negative broader impacts are well manageable.

A.2 Human Annotated Dataset

All of the 408 human-annotated summaries are randomly sampled from the summaries generated from our implemented models on arXiv and GovReport dataset. To ensure that our model summaries are annotated by human experts, we recruited three volunteers. One has years of industry experience in accounting and finance with CIMA certification while the other two are Ph.D. students of public health and computer science. Our aim for the release of the human-annotated dataset is to encourage the development of a factual consistency summarization system (model and metric). The dataset is intended for research use only. Other than that of what is already publicly available, we have taken extra steps to ensure that the factual inconsistencies generated by the summarization models do not discriminate against any individual or uniquely identify a certain person, thereby leaking information.

A.3 Model Implementation

Our model experiment in section 3 was implemented on the arXiv and GovReport with train/validation/test split of 203,037/6,436/6,440 and 17,519/974/973 respectively. Given two different pre-trained Transformers with three different input limit lengths that were tested on the baseline Longformer-only BART/PEGASUS models as well as upper-bound reduce-then-summarize models. This gives us twelve model variations per dataset. For 1K token configurations, we use BART-large and PEGASUS-large. For 4K and 8K token configurations, we follow Longformer's implementation in extending the position embedding to 4K and 8K tokens by repeatedly copying BART-large and PEGASUS-large's 1K position embeddings multiple times. All models are trained with teacher forcing on the same RTX 3090 GPU with 24 GiB of GPU memory. To save memory, we implemented gradient checkpoint. For all models with have an effective batch size of 16 where the batch size is set to be 2 and gradient accumulation step set to 8. The most expensive experiments of 8K limit require approximately 3 and 4 days respectively for Longformer-BART and Longformer-PEGASUS. As ROUGE tends to prefer longer summaries (Sun et al., 2019), we fix the maximum model output length to be 512 tokens. Generation parameters of beam search is 5 and length penalty is set to 2.0.

A.4 Factual Consistency Metric Implementation

FEQA, QUAL, FactCC, and TE-MNLI were proposed to evaluate the factual consistency of model-generated summaries under short document settings. They relied on pre-trained Transformer-based models where the input limit of 1024 tokens or lower. To extend these metric models to the long document domain, we adopt two approaches: if (i) the model requires data specific fine-tuning like FactCC, we extend the input limit of the metric model using Longformer, or (ii) the model relies on a pre-trained model that is fine-tuned on other datasets, we extract the oracle summaries of the source document where the length is the input limit of the pre-trained model.

FactCC FactCC (Kryscinski et al., 2020) implements a BERT-based factual consistency classifier that is trained on synthetic data, where the positive data labels are non-paraphrased and paraphrased sentences from the source document, and the negative labels are artificially corrupted sentences from the source document. The starting point of the BERT model is uncased, base BERT model pre-trained on English data with 512 token limits. We extend this model to 8,192 tokens using Longformer's implementation. Then, we follow the original author's work in generating the synthetic data to train our extended BERT classifier on RTX 3090 GPU with 24 GiB of GPU memory.

TE-MNLI TE-MNLI (Maynez et al., 2020) is a BERT-large classifier fine-tuned on the MultiNLI dataset (Williams et al., 2018). The classifier judged if a summary entails the document, is

neutral to the document, or contradicts the document. Multi-NLI is a sentence-level classifier. We tokenize the candidate summary into sentences and separately evaluate the factual consistency of each sentence. The score for a candidate summary equals 1 minus the average probability of contradiction for all sentences in the candidate summary. To adapt the Multi-NLI BERT-large classifier on the long document domain, we limit the total length of summary sentence and document to be less than 512 token lengths by replacing the source document with its oracle summary.

FEQA and QUAL FEQA (Durmus et al., 2020) and QUAL (Nan et al., 2021) measures factual consistency of summaries using a question-generation and question-answering (QGA) approach. This approach employs a question-generation model to generate questions from a given summary output. The generated questions are then measured in two different ways: i) answering the question conditioning on the source and ii) answering the question conditioning on the summary. If the answers match between the source and the summary, the answer is then considered consistent, otherwise, it is inconsistent. QUAL attempts to improve the efficiency of such an approach by combining the question-generation and question-answering steps into a single model. We limit source and candidate summary length to less than 512 tokens by replacing the source document with its oracle summary.

A.5 Benchmark Dataset Comparison

Long document benchmark datasets studied in this work have been used in prior research to test and compare long document summarization models. arXiv and PubMed (Cohan et al., 2018) are scientific long document summarization datasets. BigPatent (Sharma et al., 2019) is collected from U.S. patent documents. BillSum is a dataset on summarizing state bills (Kornilova and Eidelman, 2019). GovReport is a dataset of U.S. Government Accountability Office reports (Huang et al., 2021). We also compute the average result of short document datasets based on CNN-DM (Nallapati et al., 2016), NWS (Grusky et al., 2018), XSUM (Narayan et al., 2018), Reddit-TIFU (Kim et al., 2019), and WikiHow (Koupaee and Wang, 2018) in Table 5. We evaluate the document (D) and summary (S) pairs of benchmark datasets by their compression ratio, extractive coverage, extractive density and uniformity.

Compression Ratio measures the ratio of a source document length against its reference summary length. A higher compression ratio indicates larger information loss in the original document after being summarized. Compression ratios are measured based on tokens and sentences:

$$COMPRESSION_{token} = \frac{|D|}{|S|}$$

$$COMPRESSION_{sent} = \frac{||D||}{||S||}$$

Extractive Coverage and Extractive Density are introduced by Grusky et al. (2018) based on the notion of matching fragments. Fragments are obtained by greedily matching the longest shared token sequence where $\mathcal{F}(D, S)$ reflects a set of fragments with each fragment having a length represented by $|f|$. Extractive coverage calculates the percentage of tokens in summary that is a derivation of the original source text, whereas, extractive density relates to the average squared length of the extractive fragments in the summary. The former indicates the need for a model to coin novel tokens that are not in the original source text while the latter measures whether a model can match the ground truth summary merely by extracting from the original source text without rearranging or paraphrasing text.

$$COVERAGE(D, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}(D,S)} |f|$$

$$DENSITY(D, S) = \frac{1}{|S|} \sum_{f \in \mathcal{F}(D,S)} |f|^2$$

Uniformity measures whether content that are considered important by the reference summary are uniformly scattered across the entire source document. A higher score indicates that important content are scattered across the entire document with no obvious layout bias to take advantage of. This is calculated based on the normalized entropy of the decile positions of salient unigrams in the source text, where salient unigrams are the top 20 keywords extracted[3], excluding stopwords, from the reference summary.

$$UNF(unigram_{pos}) = H_{norm}(unigram_{pos})$$

[3]We use NLTK-RAKE for keywords extraction.

	Short Document Datasets					Long Document Datasets					Long vs. Short
	CNN-DM	NWS	XSum	WikiHow	Reddit	ArXiv	PubMed	BigPatent	BillSum	GovReport	Avg. Ratio
# doc-summ.	278K	955K	203K	231K	120K	215K	133K	1.34M	21.3K	19.5K	-
summ tokens	55	31	24	70	23	242	208	117	243	607	6.9x
doc tokens	774	767	438	501	444	6446	3143	3573	1686	9409	8.3x
summ sents	3.8	1.5	1	5.3	1.4	6.3	7.1	3.6	7.1	21.4	3.7x
doc sents	29	31	19	27	22	251	102	143	42	300	6.5x
Compression$_{token}$	14.8	31.7	19.7	7.2	18.4	41.2	16.6	36.3	12.2	18.7	1.4x
Compression$_{sent}$	8.3	22.4	18.9	3.3	14.5	44.3	15.6	58.7	9.7	18.1	2.2x
Coverage	0.890	0.855	0.675	0.610	0.728	0.920	0.893	0.861	0.913	0.942	1.2x
Density	3.6	9.8	1.1	1.1	1.4	3.7	5.6	2.1	6.6	7.7	1.5x
Uniformity	0.856	0.781	0.841	0.813	0.777	0.894	0.896	0.922	0.903	0.932	1.2x

Table 5: Comparison of Short and Long Document Summarization Datasets. Intrinsic characteristics are computed based on the average result of test samples. Average Ratios are computed based on the average long over short document statistics.

Fundamentals of Long Document From Table 5, the long document datasets differ from the short documents datasets in two important aspects: document length and compression ratio. Not only that long document datasets have an average document length that is 8.3 times longer than the short document datasets, they also have a considerably higher compression ratio. As compared to short documents, this suggests that either (i) there is a greater compression in the summaries, and/or (ii) the source document contains significantly more redundant information. Both aspects significantly challenge a model's ability to summarize a long document as it is required to reason over long-range dependencies.

Extractiveness and its Relationship with Compression Ratio Looking at the density value, BigPatent and arXiv are significantly less extractive than Pubmed, BillSum and GovReport. Thus, a summarizer is required to have a greater ability at paraphrasing the original document under BigPatent and arXiv. This finding is important as past work in analyzing abstractive summarization of short documents has found that the quality of model-generated summaries (Tejaswin et al., 2021; Wilber et al., 2021) and effectiveness of evaluation metrics (Gabriel et al., 2021; Pagnoni et al., 2021) to vary based on the extractiveness of benchmark datasets. Intriguingly, we further observe a strongly negative correlation, $\rho = -0.9186$, between the extractive density and the compression ratio metrics. We hypothesize that this is because, under a scenario where summary length is extremely limited, the summary writers are forced to intelligently paraphrase the source concisely so that the reference summaries can cover the salient contents.

Based on the findings above, we choose GovReport as it is the most extractive dataset with an average compression ratio, and arXiv as it is the second most abstractive dataset with the greatest compression ratio in terms of token for our systematic analysis of long document summarization systems (i.e., models and metrics).

A.6 Human Evaluation Results for Each Model Variant

Figure 5 shows human evaluation results for each model variant made in the arXiv and GovReport datasets as annotated by our volunteers.

A.7 Fine-grained analysis of Abstractive Summarizer's Factual Consistency

Figure 6 shows the types of factual errors that the abstractive models made in the arXiv and GovReport datasets as annotated by our volunteers. As a long document summary have multiple sentences and can have multiple types of errors, the total proportion may exceed 1 but the proportion of errors for each type should be lower than 1.

A.8 Human Correlation Results for Precision, Recall, F1 of ROUGE and BERTScore

Table 6 shows the correlation of ROUGE and BERTScore for precision, recall and F1 scores. When measuring relevancy of model-generated summaries, we observe F1 score to often best correlate with human judgment scores. However, when it comes to factual consistency of summaries, we do not see conclusive results as to which variant provide the best results. Furthermore, most results are not statistically significant when measuring factual consistency. Consequently, we do not include these results in our main section.

Metrics	Relevance								Factual Consistency							
	arXiv				GovReport				arXiv				GovReport			
	Pearson		Spearman		Pearson		Spearman		Pearson		Spearman		Pearson		Spearman	
	ρ	p-val	r	p-val	ρ	p-val	r	p-val	ρ	p-val	r	p-val	ρ	p-val	r	p-val
ROUGE-1																
Precision	0.08	0.26	0.07	0.30	0.26	0.00	0.30	0.00	-0.12	0.09	-0.13	0.06	**0.12**	0.10	**0.07**	0.37
Recall	0.24	0.00	0.24	0.00	0.39	0.00	0.39	0.00	**0.18**	0.00	**0.12**	0.07	-0.11	0.21	-0.15	0.12
F1	**0.29**	0.00	**0.25**	0.00	**0.53**	0.00	**0.52**	0.00	-0.08	0.26	-0.13	0.16	-0.12	0.09	-0.11	0.12
ROUGE-2																
Precision	0.01	0.92	0.02	0.83	0.27	0.00	0.36	0.00	-0.16	0.03	-0.21	0.00	**0.06**	0.44	**0.00**	0.93
Recall	**0.23**	0.00	**0.21**	0.00	0.42	0.00	0.42	0.00	**0.04**	0.52	**0.12**	0.07	-0.12	0.10	-0.11	0.15
F1	0.14	0.03	0.16	0.02	**0.43**	0.00	**0.44**	0.00	-0.12	0.09	-0.13	0.10	-0.08	0.32	-0.11	0.10
ROUGE-L																
Precision	-0.03	0.71	0.03	0.66	0.20	0.00	0.31	0.00	-0.15	0.06	-0.18	0.02	**0.12**	0.36	**0.05**	0.70
Recall	**0.22**	0.00	**0.17**	0.02	0.36	0.00	0.36	0.00	**0.09**	0.22	**0.13**	0.12	-0.12	0.09	-0.18	0.02
F1	0.12	0.07	0.17	0.06	**0.38**	0.00	**0.39**	0.00	-0.16	0.09	-0.15	0.07	-0.08	0.21	-0.11	0.11
BERTScore																
Precision	0.14	0.04	0.15	0.03	0.36	0.00	0.43	0.00	-0.13	0.00	-0.13	0.00	**0.11**	0.12	**0.05**	0.52
Recall	0.16	0.03	0.15	0.04	0.34	0.00	0.33	0.00	**0.02**	0.78	**0.03**	0.63	-0.10	0.13	-0.13	0.08
F1	**0.22**	0.00	**0.18**	0.00	0.38	0.00	0.38	0.00	-0.09	0.12	-0.10	0.10	0.00	0.95	-0.04	0.57

Table 6: Statistical Relationship between human judgement (relevance and factual consistency) and metric scores based on Pearson correlation and Spearman rank correlation coefficients and their p-values.

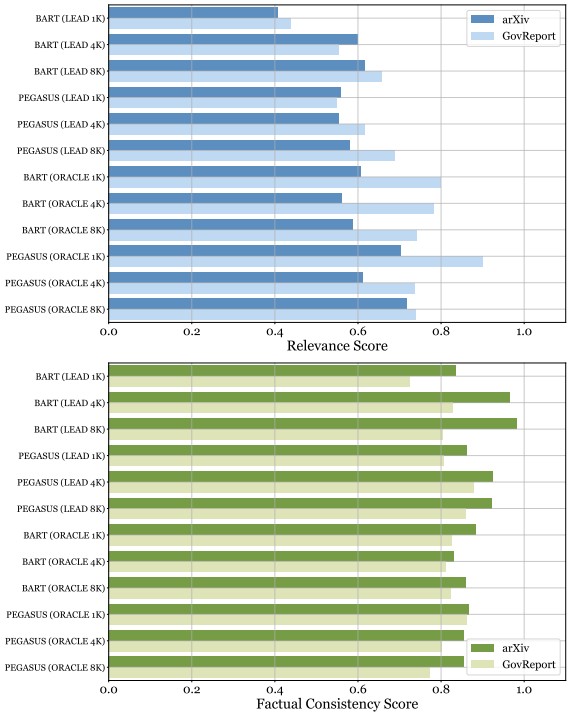

Figure 5: Human relevance (upper) and factual consistency scores for each BART model variant.

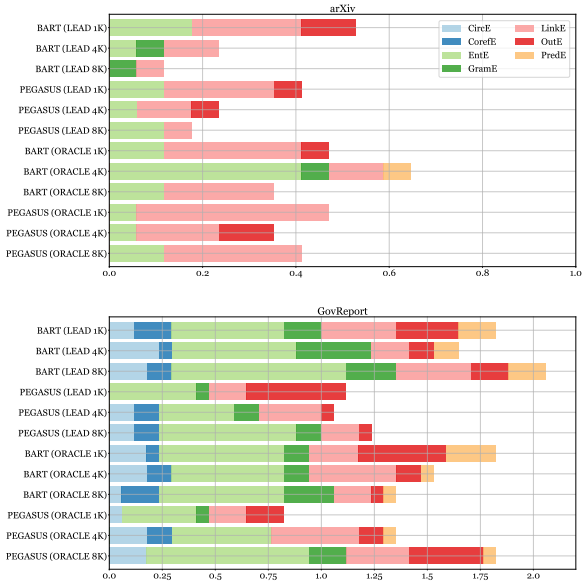

Figure 6: Factual Consistency across different model variants. The proportion for each type of error is shown based on the percentage of summaries with the same type of error. As long document summaries may have multiple sentences, each summary may have more than one type of error.