Implications of Reporting Bias within Multi-document Summarization

Emily Guo
University of California Davis
emiguo@ucdavis.edu

Erjie Zhang
University of California Davis
erjzhang@ucdavis

Abstract

Multi-document summarization (MDS) is an LLM's ability to analyze and process data from multiple documents to generate one coherent summary that fuses ideas and creates balance. However, there is a gap in our current understanding of how these requests are handled and how well these models can generate such tasks. This paper explores the problems that MDS raises in ethicality, accuracy, and bias. In our research, we evaluate several models on MDS tasks using MDS datasets. Using baseline versions of the models, we set a control for later experiments. Our results show that fine-tuning increases performance based on the metrics. We bring a level of novelty by introducing web-based LLMs into our experimentation. This crucial choice is defined by the increase in LLM usage in recent years as these summaries have the potential to affect user knowledge. We hope to showcase that analyzing MDS tasks is detailed and can be subjective to every user. We also hope to share future studies that we plan to conduct that will provide us with more insights into MDS applications.

1 Introduction

As the landscape of summarization in the context of large language models (LLMs) evolves, this change represents a new bound of researching the processing and generation of multimedia tasks. Multi-document summarization (MDS) brings new opportunities and challenges to our community. For an MDS generation task, a model will need to process multiple documents to provide one concise summary. Figure 1 pictures the basic process of MDS where each of the documents is well represented in the summary. The motivation for our research stems from the understanding that LLMs are capable of summarization. However, there is a gap in our knowledge of how well these requests can be handled.

Models can potentially spread biased information which impacts user understanding. Our research aims to bring an added level of novelty to MDS by experimentally investigating and determining model proficiency on these tasks. By addressing these topics, our work aims to contribute to the LLM landscape regarding MDS. The key focus of our research is to explore a variety of pre-trained models and LLMs and determine the effectiveness of their zero-shot and fine-tuned abilities. For our experiments, we push the need for diversity within our models' capabilities to process MDS datasets (MultiNews and Rotten Tomatoes). Experimenting with different parameters and metrics will also increase our understanding of potential biases. With this system, we evaluate the quality of the models' capabilities in producing a fair and reasonable response. In this work, we bring technical contributions to the LLM space by providing the following:

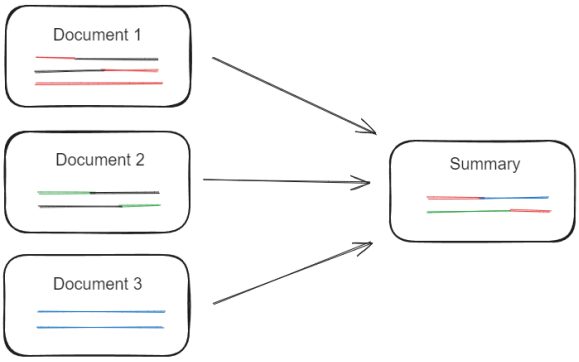

Figure 1: General Multi-document Summarization process

1. Comparative MDS experiments on various types of pre-trained models.
2. Testing involving multiple types of datasets from multiple fields of knowledge.
3. Applying MDS tasks on LLM-based Chat-bot products.
4. Analyzing and addressing issues and biases from the results.

Through this work, we aim not only to advance our current understanding of LLMs on MDS tasks but also contribute to the broader discussion on the ethical implications of generative multi-document summarization.

2 Method

2.1 Preliminaries

The project focuses on evaluating the performance of various pre-trained models and LLMs for multi-document summarization, identifying the potential biases in the generated summaries, and evaluating the quality and fairness of summarization systems across different domains and perspectives.

Such biases in the summary can be diverse. *Topic bias* is when some models perform very well in certain topics while generating low-quality summaries in some other topics. *Positional bias* (Chhabra et al., 2024) refers to a phenomenon when the model prioritizes the text in some position from the documents while ignoring some texts selectively. It has been a well-known issue in single-document summarization (Chhabra et al., 2024; Hickmann et al., 2022; Brown and Shokri, 2023). However, it is crucial to examine positional bias in multi-document summarization, given its status as a relatively new area of study. We consider this as a novel bias related to *lead bias* (Xing et al., 2021). Another bias is that some models tend to generate extractive summaries while others can summarize by abstraction. Given the same article set, different models may generate summaries with significantly different lengths, indicating a length bias. In addition, biases in ethicality are important when it comes to values and norms in decision-making by the AI models.

Based on this research topic, we have proposed the following research questions:

- What methodologies are effective in assessing reporting bias introduced by LLMs and pre-trained models in multimedia summarization?
- How accurately can current LLMs and pre-trained models capture the various nuances of multimedia content through summarization?
- How can current models of LLMs and pre-trained models be fine-tuned to represent diverse multimedia summarization better?
- How will the summative performance of LLMs pre-trained models be evaluated using the combination of metrics while considering multiple factors?

In addition, we have the following research system overview including several steps:

1. Data Collection: Collect diverse datasets containing multiple documents and corresponding human-written summaries as reference.
2. Pre-trained Models: Select and fine-tune pre-trained models, including both zero-shot and fine-tuning.
3. Evaluation Metrics and Bias detection: Define metrics to analyze the quality and biases of generated summaries from multi-document datasets.
4. Result discussions and takeaways: Analyze the results and discuss the findings from the results.

2.2 Datasets

To ensure the diversity of our data, we plan to use three different datasets covering three topics.

Our first proposed dataset is Multi-News (Fabbri et al., 2019). Multi-News is the first large-scale multi-document summarization dataset, with 56,216 article summary pairs, and each pair has multiple news articles and one reference summary written by professional editors. (Fabbri et al., 2019). It was initially introduced in 2019 and has become one of the most comprehensive datasets in the MDS field. This dataset mainly aims to test how well models can summarize social event articles.

The second dataset is the Rotten Tomatoes dataset (Wang and Ling, 2016). Rotten Tomatoes is a website that has a large scale of user reviews and critics of movies and TV shows. The Rotten Tomatoes dataset includes 3,731 movies or TV shows from the website, with both professional critics and user comments (Ma et al., 2021). Each summary from this dataset is one sentence and was created by a professional (Wang and Ling, 2016). Unlike news articles, user reviews are subjective and can be used to examine the ability of models to human opinions.

The third dataset we proposed to use is MSˆ2 (DeYoung et al., 2021). MS^2, also called Multi-Document Summarization of Medical Studies, includes over 470,000 documents and 20,000 summaries, and the authors derived these data from

the scientific literature (DeYoung et al., 2021). For non-expert users, medical documents are usually hard to read and a summary can be helpful. However, having an accurate summary is particularly necessary in this case.

2.3 pre-trained Models

In this project, we have considered using various model candidates. T5 (Raffel et al., 2023), or "Text-to-Text Transfer Transformer", is a decoder-encoder-based model widely used in natural language tasks. BART (Lewis et al., 2019) is a denoising auto-encoder by the Facebook AI group, which requires fine-tuning to adapt specific domains of tasks and improve its performance. In addition, PRIMERA (Xiao et al., 2022) is a model that was designed and pre-trained specifically for multi-document summarization (DeYoung et al., 2023). PRIMERA can be used as a high-performance zero-shot transformer as opposed to BART and T5.

Large language models(LLMs), such as ChatGPT and Gemini, are versatile and have shown their capability to understand human language inputs. It is valuable to examine MDS tasks on generative AI as it can test the capabilities of these models in summarizing diverse sets of documents while also revealing potential biases and limitations in the current products.

2.4 Metrics

In this section, we clarify the metrics in our research, covering various aspects of evaluation.

The first part encompasses the two metrics for general performance assessment, designed to detect potential biases across different textual topics and evaluate model performance. One metric is ROUGE (Lin, 2004) for lexical matching evaluation. ROUGE is the most essential evaluation method for many NLP works, such as machine translation and text summarization. ROUGE has a variation called ROUGE-L (Lin, 2004), which also helps the sentence-level evaluation and allows a more global result on the content similarity. The other one is BertScore (Zhang et al., 2020), a semantic-level metric. BertScore computes a similarity score for each token within the generated sentence and the reference sentence (Zhang et al., 2020). BertScore gives a more robust performance on challenging tasks and provides a human-like evaluation of the summary (Ma et al., 2021).

The second part is a set of data statistic metrics. Data Statistics (Grusky et al., 2020) contains multiple practical scores. One score from Data Statistics is the novelty score (Ma et al., 2021), which indicates how much new information is generated in a summary by the model. We consider a higher novelty indicates the model is good at abstractive summarization tasks but also with a higher risk of hallucination. On the other hand, with a lower novelty score, the summary can have a high redundancy and the model is more extraction-biased. Consider A is the set of original articles and s is the generated summary. and both consist of a sequence of tokens. $ST(A, s)$ computes the shared token sequences between them, which are extractive. The novelty score $N(A, s)$, or *Extractive Fragment Coverage* (Grusky et al., 2020) is calculated by (Grusky et al., 2020)

$$N(A, s) = \frac{1}{len(s)} \sum_{t \in ST(A,s)} len(t)$$

Additionally, another score given by Data Statistics is the compression ratio (Grusky et al., 2020), which computes the word count of the generated summary over the original texts (Ma et al., 2021). The compression ratio function is defined as (Grusky et al., 2020)

$$C(A, s) = \frac{len(A)}{len(s)}$$

where A is the set of original articles and s is the generated summary. The compression ratio can be used to measure how brief a summary is and help us to identify the length bias mentioned before.

To identify positional bias, we introduce a way to compute the position similarity distribution between articles. Suppose a multi-document set A has several articles. $S(a, s)$ is a similarity function that calculates the cosine similarity between the embeddings of the generated summary s and one of the original articles a. To calculate the cosine similarity, both articles and summaries have to be embedded and vectorized, where we can use a sentence transformer T. The similarity function is defined as:

$$S(a, s) = \frac{T(a) \cdot T(s)}{\|T(a)\| \, \|T(s)\|}$$

By applying this function between each article and the summary, a list of similarity scores is obtained. If one similarity between the a and s is significantly higher than others, then it implies the model has a

Metric	MultiNews	Rotten Tomatoes
ROUGE-1 Score (P/R/F1)	47.34%/23.53%/30.84%	12.86%/27.95%/17.08%
ROUGE-2 Score (P/R/F1)	13.97%/7.05%/9.20%	1.82%/4.03%/2.44%
ROUGE-L Score (P/R/F1)	24.58%/12.30%/16.08%	8.53%/19.22%/11.46%
BERT Score (P/R/F1)	84.13%/83.15%/83.63%	83.45%/85.89%/84.64%
Avg. Document Length	1717.13 words	1833.71 words
Avg. Summary Length	100.83 words	44.02 words
Compression Ratio	0.059	0.024

Table 1: Comparative Results of Vanilla PRIMERA on MultiNews and Rotten Tomatoes

positional bias on that article. We also realize different sets have distinct numbers of articles, hence a plot of graph will be helpful for us to identify positional bias.

3 Experiments

Objective

The primary objective of this study is to conduct a comparative analysis of five varying large language models (LLMs); PRIMERA, T5, Gemini, ChatGPT-4, and BART; in the context of multi-document summarization (MDS). Due to time constraints, we were only able to test with two out of the three proposed datasets. In the end, we decided to explore MultiNews and Rotten Tomatoes. Revisiting these datasets, MultiNews contains sets of news articles followed by a summary and Rotten Tomatoes contains sets of movie reviews followed by its summaries. MultiNews (MN) focuses on presenting factual content which is denser in regards to information portrayal while the documents in Rotten Tomatoes (RT) are more opinion-based which provides sentiments. The use of these two datasets allows for variety when dealing with summarization. They challenge the models by introducing a variety of summarization scenarios. Through these experiments, we explore MDS on several frontiers, enabling us to analyze a more holistic view on MDS.

In our exploration of these models using MDS, we designed three experimental approaches that would target different aspects of summarization. For the first experiment, we use zero-shot PRIMERA and T5 and evaluated these with the selected metrics: Bert-sore, Rouge-score, and Data Statistics. For both models, we generated tests for the first 100 entries from the datasets.

Vanilla PRIMERA and T5

PRIMERA is a model that is pre-trained on MDS tasks. As we are testing PRIMERA's zero-shot (vanilla) capabilities, we loaded the first 100 entries into a JSON file which we processed for generation. After PRIMERA generated 100 summaries, we were then able to assess baseline abilities. These scores, presented in Table 1, detail PRIMERA's zero-shot abilities to generate summaries from MultiNews (MN) and Rotten Tomatoes (RT). This model performs significantly better on MN than it does on RT as suggested by the higher ROUGE scores. Lower ROUGE-2 and ROUGE-L scores as compared to a moderate ROUGE-1 score denote that the model is able to perform well generating at the word level and struggles with preserving phrases at the sentence level. The model experiences this trend with both MN and RT. Alternately, BERT scores for both MN and RT are fairly equal. This denotes that there is a reasonable semantic understating that the model upholds with the human summaries. The compression ratio, provided by Data Statistics, shares the model's ability to condense the information from the documents. The low ratios indicate that the generated summaries are significantly reduced in length as compared to the original documents which allow for a more condensed and concise summary. Due to the difference in sentence structure and variation between factual and opinion heavy content, the model experiences higher performance with MN than RT. This shows that PRIMERA captures factual content degrees better than opinion-heavy documents.

We also analyze these three metrics on T5-large using the same dataset parameters. We found that, as shown in Table 2, ROUGE-1 scores are average for the MultiNews (MN) experiments with a noticeable decrease in ROUGE-2 and ROUGE-L scores. Although T5 is able to perform moderately with word level representation, larger phrases did

Metric	MultiNews	Rotten Tomatoes
ROUGE-1 Score (P/R/F1)	45.42%/16.76%/22.68%	9.37%/25.24%/13.15%
ROUGE-2 Score (P/R/F1)	11.91%/4.22%/6.11%	2.26%/6.47%/3.20%
ROUGE-L Score (P/R/F1)	26.52%/9.09%/13.28%	7.59%/20.73%/10.69%
BERT Score (P/R/F1)	83.77%/81.29%/82.49%	81.54%/84.67%/83.07%
Avg. Document Length	1717.13 words	1833.71 words
Avg. Summary Length	69.32 words	52.15 words
Compression Ratio	0.04	0.028

Table 2: Comparative Results of T5 on MultiNews and Rotten Tomatoes

Metric	MultiNews	Rotten Tomatoes
ROUGE-1 Score (P/R/F1)	55.83%/28.78%/37.40%	14.50%/29.98%/19.01%
ROUGE-2 Score (P/R/F1)	18.73%/9.78%/12.62%	3.64%/7.68%/4.80%
ROUGE-L Score (P/R/F1)	28.87%/14.86%/19.30%	10.23%/21.69%/13.49%
BERT Score (P/R/F1)	87.66%/84.63%/86.11%	83.22%/86.19%/84.67%
Avg. Document Length	1717.13 words	1833.71 words
Avg. Summary Length	108.80 words	42.98 words
Compression Ratio	0.063	0.023

Table 3: Comparative Results of Fine-tuned PRIMERA on MultiNews and Rotten Tomatoes

not transfer as effectively. However, the BERT scores are relatively on the higher end, denoting that the generated summaries retain semantic meaning from its human-generated counterpart. Similarly to the PRIMERA experiments, Rotten Tomato (RT) scores decrease significantly. Due to our lack of knowledge on the pretraining of T5-large on MDS tasks, we notice a decrease in the models ability to perform adequately on RT. The precision score for R-1 is below 10 percent while the R-2 and R-L scores are even lower. It is important to note that BERT scores for RT are just as high as the test for T5 on MN and PRIMERA. This alludes that both models have difficulties and processing and providing meaningful summaries to content that exhibits more human-like dialogue but have no issues providing summaries that retain semantics; as suggested by BERT scores.

Therefore, T5 is able to process subjective data far more efficiently than opinion-based data. The lower ROUGE scores compared to PRIMERA note that PRIMERA's pretraining on MDS tasks equates to a slight improvement of generative summarization. Overall, PRIMERA and T5 are able to perform well on MulitNews but has difficulty with Rotten Tomatoes. This first experiment shares the baseline of what models are capable of when it comes to MDS tasks. This sets the stage for the next experiments that will further test model performance on MDS.

Fine-tuned PRIMERA

Experimenting with zero-shot models alone is not enough to bring novelty in MDS. We introduce PRIMERA, fine-tuned on MultiNews, to experiment with summarization performance. Keeping the parameters static, we use BERT-score, ROUGE-score, and Data Statistics for the metrics as well as testing this model on MultiNews and Rotten Tomatoes. The goal is to perform comparative analysis on fine-tuned versus zero-shot models to determine the degree that training affects a models ability on MDS tasks.

When testing with the MultiNews dataset (MN), the fine-tuned version of PRIMERA showcases improvement as compared to its vanilla counterpart, as seen in Table 3. This change is most apparent in the ROUGE scores. R-1 improved by 8.49%, R-2 by 4.76%, and R-L by 4.29%. The increase in precision and recall account that fine-tuning allows a model to capture word, phrase, and sentence representation more effectively. The BERT score increase from 84.13% to 87.66% which is a 3.53% increase in precision. This improvement shows that the model is able to represent semantics from the original documents and reflect them in the generate summaries.

Interestingly enough, PRIMERA did not show much improvement on the Rotten Tomatoes dataset. With an average of 1.72% increase across all

Metric	MultiNews	Rotten Tomatoes
ROUGE-1 Score (P/R/F1)	48.38%/20.94%/28.30%	8.70%/41.17%/13.61%
ROUGE-2 Score (P/R/F1)	15.17%/6.71%/9.18%	3.30%/12.05%/5.00%
ROUGE-L Score (P/R/F1)	29.00%/11.59%/15.87%	6.80%/34.01%/10.83%
BERT Score (P/R/F1)	86.92%/82.68%/84.74%	82.11%/86.76%/84.37%
Avg. Document Length	1531.7 words	1899.7 words
Avg. Summary Length	81.5 words	109.8 words
Compression Ratio	0.053	0.058

Table 4: Comparative Results of Gemini on MultiNews and Rotten Tomatoes

Metric	MultiNews	Rotten Tomatoes
ROUGE-1 Score (P/R/F1)	38.27%/43.83%/40.33%	4.41%/57.21%/8.15%
ROUGE-2 Score (P/R/F1)	10.45%/1.22%/11.14%	1.53%/19.81%/2.83%
ROUGE-L Score (P/R/F1)	17.00%/19.80%/18.15%	3.52%/46.09%/6.51%
BERT Score (P/R/F1)	85.59%/84.65%/85.11%	81.11%/87.28%/84.08%
Avg. Document Length	1531.7 words	1899.7 words
Avg. Summary Length	235.8 words	253.4 words
Compression Ratio	0.154	0.133

Table 5: Comparative Results of ChatGPT-4 on MultiNews and Rotten Tomatoes

ROUGE scores and a 0.23% decrease in ROUGE precision, fine-tuning the model on MultiNews did not improve the models ability to summarize opinion-based content. Overall, this experiment details PRIMERA'S sensitivity to various types of data shown through the increase in precision across all fronts when it comes to MultiNews but the minimal improvement in Rotten Tomatoes.

Gemini and ChatGPT-4

The third experiment focuses on testing web-based products to assess the capabilities of highly accessible LLMs in MDS tasks. In the previous experiments, we explored MDs capabilities with pre-trained and fine-tuned models. However, this alone does not innately explore our method to its fullest. Due to the accessibility of LLMs like ChatGPT-4 and Gemini, it is important to understand the level of responsibility that these companies withhold. As the use of LLMs increases, users are relying on their interactions with these models more than ever. The purpose of this experiment is to determine the ability for these LLMs to relay concise and accurate summaries.

For the setup of this experiment, we use Gemini and ChatGPT-4 as these are competitors to each other. Because this is a manual experiment, we limit the data to the first 10 entries and their summaries.

As seen in Table 4, Gemini exhibits moderate performance in ROUGE-1 but lower scores in ROUGE-2 and ROUGE-L. The LLM is capable of capture key words from the documents but not at the phrase and sentence level. However, the high BERT score indicates the LLM's ability of maintaining semantic understanding. As for Data Statistics, the model has a moderate compression ratio; meaning that the model balances its summary generation with that of the human summaries from the dataset. When comparing this data with the previous tests with the pre-trained models, there is not a significant increase in any of the scores.

ChatGPT, being the most universally well-known, is an important model to explore. For these experiments, we orchestrated the same setup with the results shown in Table 5. The results are shocking. Not only does ChatGPT-4 have mediocre scores, this LLM also exhibits the lowers ROUGE scores across all three experiments. This may be due to ChatGPT's preference to produce original content. With many users needing new and clever responses, this may impact the LLM's ability to retain similar structures within their summaries.

When comparing Gemini and ChatGPT-4, Gemini tends to have less lengthy summaries as compared to GPT's. They both also experience difficulties with summarizing Rotten Tomtaoes datasets. This is a prominent challenge that all models have experienced throughout each trail. However, assessing MDS tasks is not as simple as using these

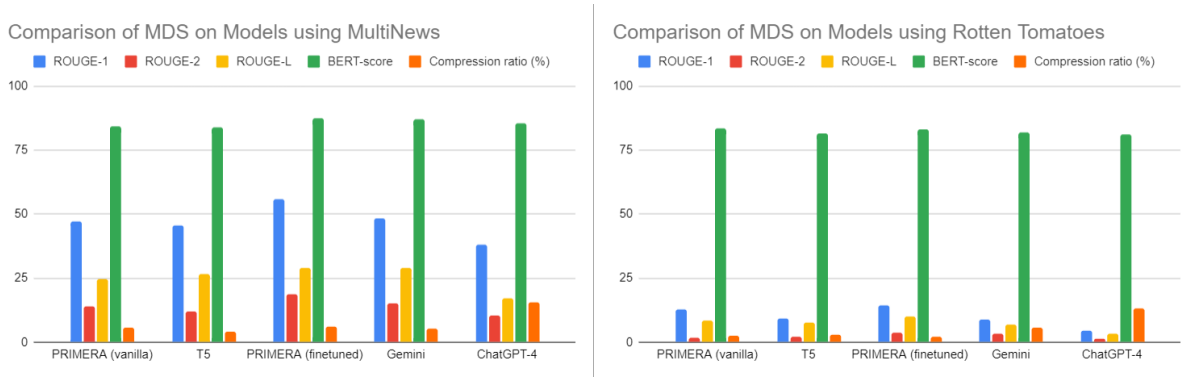

Figure 2: Graph depicting all results

three metrics. It is important to consider that human evaluation has equal weights. Although we did not have time to survey these summaries, we understand that summarization ability is not constricted to technical metrics.

3.1 Takeaways and Discussion

By comparing the results on both Multi-News and Rotten Tomatoes datasets horizontally, we have noticed that these models are generally better at summarizing news articles, not movie reviews. For all of these models, their benchmark scores on Rotten Tomatoes are much lower than the ones on Multi-News. This implies that it is much more challenging to summarize human opinions, which are generally more controversial and diverse. Therefore, we may consider these models to have a topic bias on tasks. Refer to Figure 2 to see this comparison.

For length bias, PRIMERA and T5 do not show much difference in the compression ratios. However, ChatGPT-4 and Gemini are generating much more lengthy summaries compared to PRIMERA and T5. This is particularly noticeable when generative AIs summarize Rotten Tomato reviews. Even though each movie review and human-written summary is about one sentence in length, LLMs are still generating long paragraphs, sometimes with bullet points. This might be because we did not prompt the models about how long is summary is during the experiment.

As PRIMERA is a domain-specific model for MDS tasks, it outperforms T5 as we expected. However, based on reading some summaries of these two models, we have realized they both have limitations in the coherence and fluency of context. PRIMERA tends to extract the exact texts from the original sets of documents[1], and T5 seems to be extractive as well, but with a lower readability. On the other hand, fined-tuned PRIMERA generates more abstractive summaries with better readability and fluency. This is consistent with our scores on metrics because of the large leap in both ROUGE and BertScore after testing fine-tuned PRIMERA. We consider PRIMERA(vanilla) and T5 to be more extraction-biased, while PRIMERA(fine-tuned) is more abstraction-biased. For ChatGPT-4 and Gemini, generative AIs basically never extract the exact sentences from the documents, hence it is trivial that they have strong abstractive bias.

In general, MDS is still a very challenging topic even for generative AIs. Both ChatGPT-4 and Gemini have some different levels of difficulties in summarizing lengthy multi-documents. One interesting part during the test of the Gemini web version is that Gemini failed to answer two sets of new articles. For example, the output from Gemini is "*I'm still learning how to answer this question. In the meantime, try Google Search.*" and "*If you'd like up-to-date information, try using Google Search.*" Further research is necessary to analyze the current limitations of LLMs.

4 Related Work

How "Multi" is Multi-Document Summarization? (Wolhandler et al., 2022), published to EMNLP in 2022, presents an argument that pushes the need to measure how "disperse" generated summaries are compared to the document summaries from the datasets. When proposing their own measure, the authors note the importance of assessing

[1]When testing both PRIMERA(vanilla) and PRIMERA(fine-tuned), the default settings from the authors limit the number of new tokens to 178. This ends up by cutting off the last sentence from PRIMERA, and the generated summaries are incomplete.

MDS tasks this way. In our research, we focus on the general capabilities of LLMs to perform MDS on certain datasets. This paper introduces an added level of analysis that we can perform during our experiments to test not only the base-level of performance but also how efficient data is parsed and grouped to create a summary. There are multiple methods to approach when discussing MDS tasks. These two studies have the potential to motivate each other as one provides novel metrics and the other, novel experiments and analysis.

In **A Hierarchical Encoding-Decoding Scheme for Abstractive Multi-document Summarization** (Shen et al., 2023), a "Findings" paper from EMNLP 2023, the authors note that concatenated source documents are used in pre-trained models or new MDS models from previous works, while the complicated but unique relationships between multiple documents are not utilized. The authors propose a novel method based on hierarchical encoding-decoding (HED) that handles the features of cross-document information and improves the general performance of MDS models. pre-trained models such as PRIMERA, BART, and LongT5, are tested by 10 different datasets, including Multi-News and Rotten Tomatoes.

An EMNLP 2020 Findings paper **Corpora Evaluation and System Bias Detection in Multi-document Summarization** (Dey et al., 2020) introduces a clear definition for the standard of multi-document summarization. Authors list various metrics for evaluating the MDS corpus, such as ROUGE, abstractness, inter-document similarity, redundancy, etc. The authors also perform an evaluation on the biases that existed in the current MDS systems affected by the properties of documents.

5 Conclusion and Future Work

In this paper, we highlight the importance of understanding current LLM and pre-trained model capabilities on MDS tasks. With the landscape of technology revolving around natural language possessing and large language models, it is crucial to understand why these models affect user knowledge. We explore several models, PRIMERA, T5, Gemini, and T5 along with evaluation metrics: BERT-score, ROUGE-score, and Data Statistics. We ultimately chose MultiNews and Rotten Tomatoes as our datasets as they vary in data content and structure. The first method of experimentation tested baseline model capabilities. In this way, we were able to determine the quality of PRIMERA and T5's generation. To our understanding, using base-lines is not enough. As this is our control, we experimented with finetuned models. Using PRIMERA which was finetuned on MultiNews, we tested its ability to complete MDS tasks on both datasets. The experiments showed an increase of relevancy and structure in the generated summaries as compared to the vanilla generations. Additionally, we access MDS on highly accessible LLMs. Due to the sheer use of these web-based models, it is important to compare these LLMs to the previous models. From the results, we are not able to see much difference in scores. However, it is crucial to understand that analyzing these models' capabilities on MDS extends further on using metrics like ROUGE-score, BERT-score, and Data Statistics. There are several other methods to be considered. The effectiveness of a model to capture multiple documents is subjective to every user. In our future work, we plan to explore more of these methods.

5.1 Future Work

Due to constraints in time and resources, our future work will mainly focus on completing the following experiment plan:

- The entire test set will be involved during the testing, instead of the first 100 samples. Sufficient test samples will help eliminate the potential risks during our testing. In addition, granting API access from Google and OpenAI is necessary to automate the experiments.
- We will finalize fine-tuning BART and T5, as well as all models on Rotten Tomatoes and MSˆ2. Fine-tuning requires GPU access and time, and it is important to evaluate their effectiveness in summarizing human opinions after adopting this specific domain with training data.
- We will finish the implementation of metrics for bias detection. This includes the proposed methods in the Metrics section, such as novelty score in Data Statistics and the Positional similarity distribution model. This is essential to systematically detect extraction or abstraction bias and positional bias in a summary of multiple documents.
- In addition, a more comprehensive and systematic human evaluation is required to review the quality of the summary produced by

models. We need to set up an evaluation protocol and collect user feedback on the fluency, readability, and relevancy of the summaries.

References

Hannah Brown and Reza Shokri. 2023. How (un)fair is text summarization?

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias.

Alvin Dey, Tanya Chowdhury, Yash Kumar Atri, and Tanmoy Chakraborty. 2020. Corpora evaluation and system bias detection in multi-document summarization.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies.

Jay DeYoung, Stephanie C. Martinez, Iain J. Marshall, and Byron C. Wallace. 2023. Do multi-document summarization models synthesize?

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2020. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies.

M. Lautaro Hickmann, Fabian Wurzberger, Megi Hoxhalli, Arne Lochner, Jessica Töllich, and Ansgar Scherp. 2022. Analysis of graphsum's attention weights to improve the explainability of multi-document summarization.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2021. Multi-document summarization via deep learning techniques: A survey.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. A hierarchical encoding-decoding scheme for abstractive multi-document summarization.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Ruben Wolhandler, Arie Cattan, Ori Ernst, and Ido Dagan. 2022. How "multi" is multi-document summarization?

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. Primera: Pyramid-based masked sentence pre-training for multi-document summarization.

Linzi Xing, Wen Xiao, and Giuseppe Carenini. 2021. Demoting the lead bias in news summarization via alternating adversarial learning.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.