Corpora Evaluation and System Bias Detection in Multi-document Summarization

Impact & Alternatives

Yannick Lang
Study Program: B.Sc. Applied Computer Science

University of Bamberg, Bamberg, Germany
yannick-stephan.lang@stud.uni-bamberg.de

Abstract. Large language models are on the rise and as such have an increasing contact surface with more and more people. Partly because they are trained on data produced by people, systems that generate language may reproduce or even insert bias. This paper focuses on metrics for quantifying bias in multi-document summarization corpora and systems, commonly used to compress news articles or similar area where multiple documents deal with a common topic. Different proposals for such metrics will be discussed and their usage assessed. In the end, the "Wikipedia Current Events Portal" corpus will be benchmarked on selected metrics.

Keywords: Bias · Summarization · Corpora · Language Models

1 Introduction

With the recent rise in popularity of large language models such as GPT3[**?**] or DALL-E[**?**], and companies beginning to bridge the gap between research and commercial use, it is important to keep in mind that such models may not be free of biases.[**?**]

This paper looks at approaches to quantify bias in both actual systems and underlying corpora of multi-document summarization (MDS) models. MDS has use cases such as review or news aggregation, and as such biases can have noticeable impact.[**?**]

Starting from a 2020 paper[**?**] dealing with this topic, the metrics for quantifying biases proposed in this paper as well as its impact and possible alternative criteria will be discussed. Finally, the metrics will be applied to a recently published corpus, for which they where not provided at publication and possible conclusions about its structure will be discussed.

2 Original Paper

The starting point and primary source for this paper will be "Corpora Evaluation and System Bias Detection in Multi-document Summarization"[**?**], hereafter referenced as "original paper" or "Dey et al.", which was published in late 2020. In

it, the authors propose several metrics for quantifying biases in corpora and systems for multi-document summarization models, enabling a better comparison between them.

To showcase their metrics, they apply them to several high profile corpora, analyze the results and come to the conclusion, that researchers should consider including those metrics when publishing new corpora in order to facilitate comparisons.

2.1 Metrics

The paper proposes a list of metrics which can be used to gauge certain attributes of a corpus. These metrics are:

- **Inter Document Similarity (IDS)**, which shows the similarity between each documents
- **Pyramid Score**, "defined as the ratio of a reference summary score and an optimal summary score"[**?**], i.e. how good the reference summaries are
- **Inverse Pyramid Score**, which measures the influence of documents on a the reference summaries
- **Redundancy** describes the density of information in the documents

For MDS systems, they also present a list of metrics aimed at evaluating the performance and properties of a given MDS systems:

- **ROUGE** one of the most basic metrics in text summarization, which has been used to measure similarity between generated summaries and the references included in the dataset via recall since 2004[**?**]
- **F1 Score** - similar to ROUGE but considers both recall and precision
- **Inter Document Distribution** - similarly to the Inverse Pyramid Score metric for corpora, this measures the influence of each document on the generated summary
- **Redundancy** describes the summaries coverage of information from the documents

Additionally, some metrics are suitable for both both corpus and system, where the reference summary is used for corpus evaluation and the generated summary for system evaluation.

- **Abstractness** quantifies the similarity between the generated or reference summary and the associated documents, where less similarity means higher abstractness
- **Layout Bias** measures the distribution of information within a document for corpora and the distribution of sections in the documents that provide the information in the generated summary for systems

The authors categorize those metrics into subjective and objective. They assign the highest importance to the objective metrics of Pyramid Score and Inverse Pyramid Score, and propose that scientists introducing new MDS corpora shall at least report values for those, although ideally all metrics should be considered.

2.2 Results

The authors calculate their proposed metrics for a range of relevant MDS corpora and systems. Their main finding is a high influence of the corpus used on the performance of a given system.

For corpora, the authors highlight differences between datasets that were quantified by their metrics. They also discern trends over time and between crowd-sourced and non crowd-sourced datasets.

Three of the five corpora analyzed in the paper consist of news articles, which make up a lot of (particularly early, i.e. early 2000) MDS corpora. This prominence can be explained by the availability of news articles as well as the value of the real world use case of combining and summarizing news articles.

This explains the papers focus on layout bias, i.e. the influence a tokens position in the document has on the likelihood of making it into the summary. Layout Bias can be especially prominent in news articles, since reporters aim to provide readers with an overview of the topic within the first few sentences. This leads to a higher amount of information relevant to the summary in these early sentences. The authors show empirically that layout bias present in a corpus carries over to trained models, i.e. that the system will then be more likely to include information from earlier sentences, even when the input does not show the same layout bias as the training material.[**?**]

A further key observation from the paper is that abstractness in the training corpus is correlated with the abstractness of system generated content.

The influence of corpus on systems can be seen in 1, which ranks system performance as measured by ROUGE-1/ROUGE-2/F1 Score respectively for every corpus under consideration. On each dataset, a different MDS system performed best, as denoted by rank 1 and highlighted by the blue color.

For ROUGE-1, the only outlier is ICSI_Summ, which managed to be the best-performing system on two separate corpora, but also achieved the worst scores on the DUC corpus and the third-worst on Opin. This indicates the reliance on specific datasets, as opposed to high performance independent of input.[**?**]

The other collected metrics performance metrics, such as ROUGE-2 or F1 Score, do not show the exact same, but still fundamentally similar patterns. When using F1 to judge quality of generated summaries, for example, there are two systems that achieve the highest score on two separate Corpora instead of just one. The main observation, that no system strictly outperforms all others, still holds.

Therefore, to make reliable statements about the capabilities of a MDS system, it is not sufficient to just choose one dataset and report ROUGE scores for that. Instead, performance over multiple datasets should be considered and reported.

2.3 Key Takeaways

The authors state that a lack of clear definition for MDS tasks leads to a lack of a single (or a few) standardized dataset. Instead, most scientists provide their

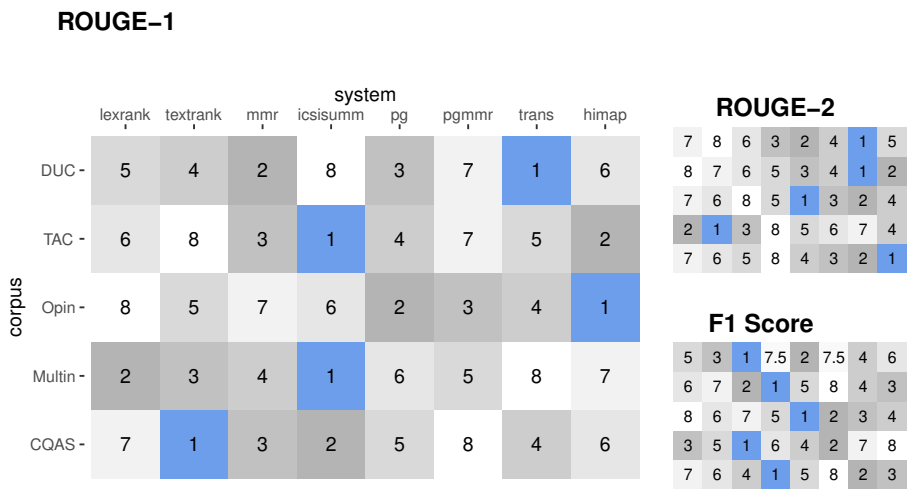

Fig. 1. Overview based on the results from the initial paper, where for every MDS system performance over each dataset is ranked and the best performing system for every dataset is highlighted. Lower number indicates better score.

own custom dataset with newly proposed MDS systems. This severely reduces the comparability between systems for two reasons.

For one, the authors have established the influence a specific dataset can have on the performance of MDS system (see 2.2). Therefore, the reported performance of a newly proposed system could be inflated when it is only evaluated on a custom corpus as compared to established corpora. However, using an existing corpus may not always be viable, due to the unstandardized nature of MDS tasks. To increase comparability between datasets, the authors thus aim to establish metrics for comparing datasets. In order to be useful, these metrics should be reported by researchers creating a new corpus. Objective metrics include Pyramid and Inverse Pyramid Scores and are assigned a higher importance by the authors, who state those two "must be reported as they are strong indicators of generic corpus quality"[**?**].

Regarding biases, the original paper focuses mostly on structural bias as opposed to in the content. Due to layout bias, a system trained on a news corpus may now perform well on other corpora with documents following a similar structure, but also significantly worse on collections of documents that either have no layout bias at all or a different, not as front-loaded, kind of layout bias.[**?**] The authors establish a connection between bias in training corpus and system predictions, but point out the need for future research regarding explicit causes and potential counter measures.

2.4 Impact

Now that the benefits and use cases of universally accepted and reported metrics for MDS corpus and system attributes haven been established, it is time to look at the impact of the paper. Did the other scientists build upon these proposed metrics or apply them to their own system or corpus?

To answer this question, forward search was performed via the online research tool SemanticScholar.org. This reveals publications that reference "Corpora Evaluation and System Bias detection in Multi Document Summarization", which allows an assessment of its impact in the scientific world.

During the almost two years since its publication in October 2020, the paper was referenced 5 times. SemanticScholar.org classifies those citations into three categories: [**?**]

- Background - 3 citations
- Methods - 1 citation
- Results - 1 citation

It also identifies one citation, in "ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining" (ConvoSumm)[**?**], as "highly influential"[**?**]. ConvoSumm propose a benchmark of summaries and performances for four "widely-used datasets". The authors use the metrics for Inter-document Similarity, Redundancy and Layout Bias proposed in the original paper to evaluate their data. They also publish the results and their key finding from applying the metrics, which is a reduced layout bias in their data compared with the mostly news focused datasets that were used in the original paper. While they do not use every metric, their choice seems reasonable, but deviates from the original papers authors assessment of the metrics importance, as Dey et al. consider Pyramid Score and Inverse Pyramid score the most important metrics to report.[**?**]

The second citation is in a recently published paper called "Two-phase Multi-document Event Summarization on Core Event Graphs" [**?**]. The authors also uses a selection of the proposed metrics to evaluate their datasets, namely abstractness and IDS.

The other three papers reference the original paper in a minor way, usually to back up general statements or disclaimers about MDS corpora potentially containing bias or the possibility of bias propagating to the generated content.[**?**,**?**,**?**]

An overview over the types of citations is given in 1.

In summary, it can be stated that the original paper has not yet had a lot of impact, which is not all that surprising given its somewhat recent publishing at the end of 2020. Two of the five studies that reference this paper actually use the proposed metrics, however it stands out that neither of those use the two metrics the original authors deemed most important.

3 Alternative Criteria & Status Quo

In the last section we determined that, as of August 2022, no paper has built upon these metrics, either proposing changes to them, criticizing or extending

uses metrics

	IDS	Abstractness	Layout Bias	Redundancy
ConvoSumm [**?**]	x		x	x
2-Phase [**?**]	x	x		

generic citation

	citation for:
Segmentation [**?**]	existence of bias
What is a Summary? [**?**]	existence of bias
Counseling [**?**]	past studies

Table 1. Breakdown of the context in which the five papers reference the original paper. In the left table, only metrics that were used are displayed, the x character indicates that this paper used this metric.

them. We will now show other metrics and approaches to bias employed by researchers that (recently) published MDS corpora or systems.

At the beginning of their paper, Dey et al. state the problem of missing standard corpora due the non-standardized tasks of MDS. The same is currently true for metrics of bias. While the authors aimed to establish a framework for this, the previous section showed that they have not yet succeeded.

As a result, some papers use different metrics, that may aim at similar attributes. For example, a 2016 paper proposing a new MDS corpus uses textual heterogeneity as a metric. [**?**] While not exactly the same, it is introduced to determine the similarity of documents, serving the same purpose as IDS in the original paper.

Similarly, for the "Wikipedia Current Events Portal" (WCEP) dataset, which will be analyzed more in-depth later in this paper, the authors used "coverage" and "density" metrics to quantify abstractness, similarly to the Abstractness metric.[**?**] Coverage and density have been used for single document summarizations earlier.[**?**]

2 shows a selection of recent papers introducing MDS corpora and their approach to bias and comparability.

	mention of bias	comparability metrics
WCEP[**?**]	no	coverage, density
MS2[**?**]	no	-
HowSumm[**?**]	no	coverage density
QMSumm[**?**]	mentioned, but not quantified	-

Table 2. Recent papers introducing MDS corpora and their handling of bias/comparability. All papers also report basic statistics, i.e. metrics such as count and average length and performance of select MDS systems.

While Dey et al. focus on structural biases, especially Layout Bias, corpora generated from human content, i.e. most if not all corpora, also display biases in

their content. These usually carry over through MDS systems into the generated summaries. [?]

Even if the training corpus does not contain biases, systems have been shown to introduce biases, such as hallucinations, i.e. statements in the generated summary that are not included in the candidate documents and do not follow from them. This is correlated with exposure bias, caused by differences in documents between training corpus and data used for actual work. [?]

4 Own Analysis

For this paper, I will calculate the metrics Dey et al. proposed for MDS corpora to a dataset myself. The corpus was selected on the following criteria:

- must be MDS corpus - the metrics were specifically designed for MDS corpora, and some, like Inter Document Similarity, only make sense in this context.
- must not already have been analyzed with these metrics, either in the original paper or by its creators
- should be somewhat recent
- dataset must be publicly available

This last criterion was actually the most constraining. I opted for the "Wikipedia Current Events Portal" (WCEP) dataset, proposed in 2020. [?] It meets all of the conditions mentioned above, as it was published in 2020, was designed explicitly for MDS, does not include the metrics from the original paper (which was published after WCEP) and the authors provide a public download link to the entire dataset on their GitHub.[?]

Although the original papers authors state that they "develop an interactive web portal for imminent corpora to be uploaded and evaluated based on [their] proposed metrics", I was unable to find any link or other kind of reference to this. The paper does, however, include a link to the source code for their analysis on GitHub, which although lacking documentation, served as the starting point for the following analysis of WCEP.

For this paper I have forked the repository and updated the code for the corpus metrics to work with the WCEP dataset. I have also added documentation on how to use the code and which format the data is expected to have, which is missing in the original version.

The WCEP dataset is split into 3 parts, a vary large training set and the test and validation sets, which each make up about 10 percent of the total amount. For this evaluation, I treated them as separate datasets, so that I could determine if they are homogenous.

For this task, I modified the source code provided by the authors slightly, in order to make this analysis possible. The version of the code used for this analysis is available on my Github [1].

[1] Repository: https://github.com/layaxx/summarization_bias

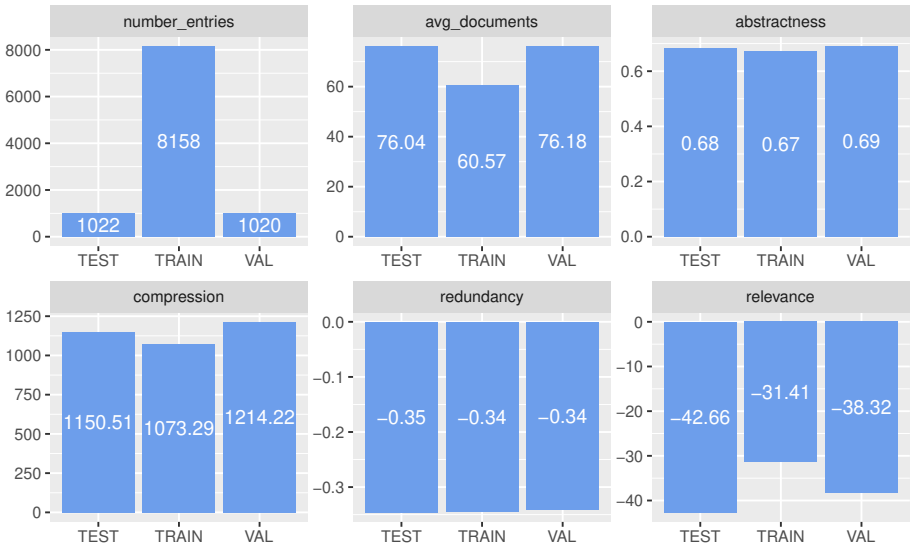

Fig. 2. Overview over the benchmark results calculated from the WCEP corpus. Every metric is reported for each of the testing, training and validation subsets of WCEP.

The results can be seen in 2. "number_entries" denotes the number of topics in each set, consisting of one reference summary and a set of documents. The average number of documents per set is shown as "avg_documents". "abstractness", "compression", and "redundancy" are the same as described in the original paper. With the "relevance" metric, I am unsure whether this is identical to IDS or just a precursor to its calculation.

While the other metrics were calculated for the complete set, due to high computational demands redundancy and relevance have been calculated on a random sample consisting of 100 topics for each of the three subsets.

I have not been able to measure the Layout Bias and the Pyramid Scores, as it is unclear to me what format the original authors code expect for the input to calculation.

Apart from obvious differences in size, the subsets of WCEP seem pretty homogenous after analysis. Notable differences are, however, visible on the relevance and average documents per topic metrics. The training set contains, on average, 15.47 (20.34%) less documents per topic compared with the test set and 15.61 (20.49%) less than the validation set. This can be seen well in 5, where the training subset contains significantly less frequently topics with 100 documents. The relevancy of the training set is also notably closer to 0 compared with the other subsets.

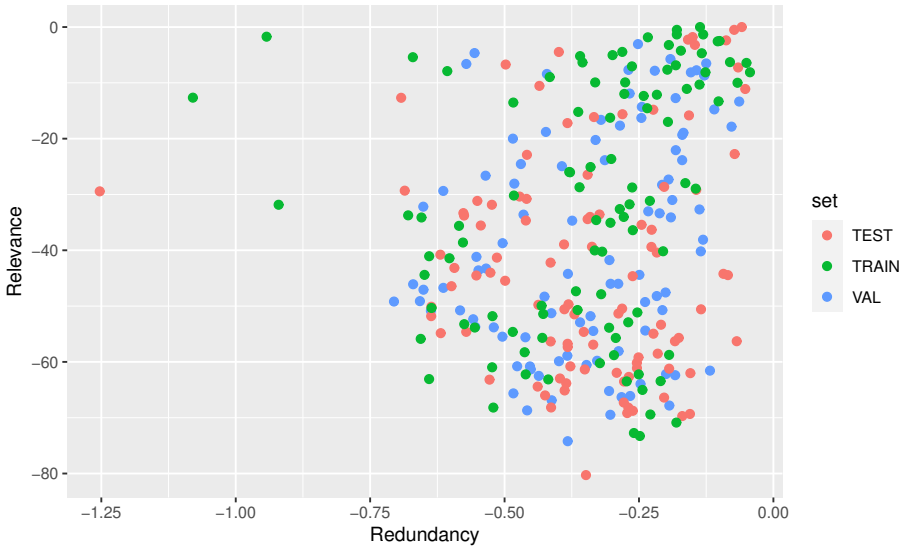

Fig. 3. Scatter plot of Relevance and Redundancy shows no obvious difference between the subsets.

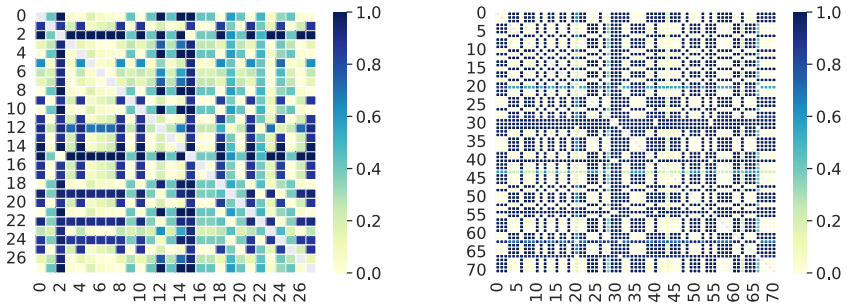

Fig. 4. Example plot of IDS for topic 70168 from the test subset and topic 62667 from the training set.

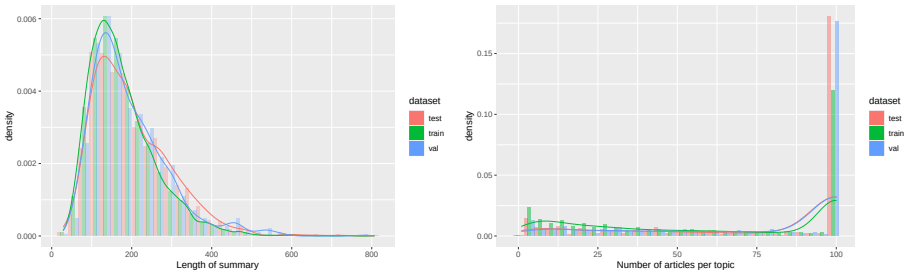

Fig. 5. Distributions of summary length and number of articles per topic in the subsets. Left graph showcases subtle differences in length of summary distribution, right graph shows the cause for the difference in average number of articles discussed previously

5 Summary

This paper summarized possible metrics for measuring and comparing attributes of multi-document summarization corpora and systems, as proposed in the original paper. [**?**]

Based on these, possible alternatives were discussed and the status quo of metrics in MDS was outlined. The metrics were then calculated for the WCEP corpus.

Summarizing the above, the field of MDS is often not defined very clearly and therefore missing standard. This is also true for dealing with biases, which are, if at all, mostly discussed qualitatively as a sort of disclaimer, but not usually quantified. Future Research into standardizing corpora and metrics for MDS may be helpful to ensure a productive debate about bias in generated language and possible approaches to mitigating them.

References

1. Adams, G., Alsentzer, E., Ketenci, M., Zucker, J.E., Elhadad, N.: What's in a summary? laying the groundwork for advances in hospital-course summarization. Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting **2021**, 4794–4811 (2021)
2. Boni, O., Feigenblat, G., Lev, G., Shmueli-Scheuer, M., Sznajder, B., Konopnicki, D.: Howsumm: A multi-document summarization dataset derived from wikihow articles (2021). https://doi.org/10.48550/ARXIV.2110.03179, https://arxiv.org/abs/2110.03179
3. Chen, Z., Xu, J., Liao, M., Xue, T., He, K.: Two-phase multi-document event summarization on core event graphs. J. Artif. Intell. Res. **74**, 1037–1057 (2022)
4. Cohan, A., Ammar, W., van Zuylen, M., Cady, F.: Structural scaffolds for citation intent classification in scientific publications. In: NAACL (2019)
5. complementizer: complementizer/wcep-mds-dataset (Feb 2022), https://github.com/complementizer/wcep-mds-dataset

6. Dey, A., Chowdhury, T., Kumar, Y., Chakraborty, T.: Corpora evaluation and system bias detection in multi-document summarization. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2830–2840. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.254, https://aclanthology.org/2020.findings-emnlp.254
7. DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B., Wang, L.L.: Ms2: Multi-document summarization of medical studies (2021). https://doi.org/10.48550/ARXIV.2104.06486, https://arxiv.org/abs/2104.06486
8. Fabbri, A.R., Rahman, F., Rizvi, I., Wang, B., Li, H., Mehdad, Y., Radev, D.: Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In: ACL (2021)
9. Gholipour Ghalandari, D., Hokamp, C., Pham, N.T., Glover, J., Ifrim, G.: A large-scale multi-document summarization dataset from the Wikipedia current events portal. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1302–1308. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.120, https://aclanthology.org/2020.acl-main.120
10. Grusky, M., Naaman, M., Artzi, Y.: 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). pp. 708–719. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1065, https://aclanthology.org/N18-1065
11. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
12. Moro, G., Ragazzi, L.: Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In: AAAI (2022)
13. Nadeem, M., Bethke, A., Reddy, S.: StereoSet: Measuring stereotypical bias in 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 (Volume 1: Long Papers). pp. 5356–5371. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.416, https://aclanthology.org/2021.acl-long.416
14. OpenAI: Gpt-3 powers the next generation of apps (Mar 2021), https://openai.com/blog/gpt-3-apps/
15. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022). https://doi.org/10.48550/ARXIV.2204.06125, https://arxiv.org/abs/2204.06125
16. Srivastava, A., Suresh, T., Lord, S.P., Akhtar, M.S., Chakraborty, T.: Counseling summarization using mental health knowledge guided utterance filtering. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 3920–3930. KDD '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3534678.3539187, https://doi.org/10.1145/3534678.3539187
17. Tamkin, A., Brundage, M., Clark, J., Ganguli, D.: Understanding the capabilities, limitations, and societal impact of large language models (2021). https://doi.org/10.48550/ARXIV.2102.02503, https://arxiv.org/abs/2102.02503

18. Valenzuela, M., Ha, V.A., Etzioni, O.: Identifying meaningful citations. In: AAAI Workshop: Scholarly Big Data (2015)
19. Wang, C., Sennrich, R.: On exposure bias, hallucination and domain shift in neural machine translation. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.326, https://aclanthology.org/2020.acl-main.326/
20. Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Awadallah, A.H., Celikyilmaz, A., Liu, Y., Qiu, X., Radev, D.: Qmsum: A new benchmark for query-based multi-domain meeting summarization (2021). https://doi.org/10.48550/ARXIV.2104.05938, https://arxiv.org/abs/2104.05938
21. Zhu, C., Yang, Z., Gmyr, R., Zeng, M., Huang, X.: Leveraging lead bias for zero-shot abstractive news summarization (2019). https://doi.org/10.48550/ARXIV.1912.11602, https://arxiv.org/abs/1912.11602
22. Zopf, M., Peyrard, M., Eckle-Kohler, J.: The next step for multi-document summarization: A heterogeneous multi-genre corpus built with a novel construction approach. ACL Anthology p. 1535–1545 (Dec 2016), https://aclanthology.org/C16-1145/