Pathology Dynamics at BioLaySumm: the trade-off between Readability, Relevance, and Factuality in Lay Summarization

Irfan Al-Hussaini, Austin X. Wu, Cassie S. Mitchell
Georgia Institute of Technology
alhussaini.irfan@gatech.edu, awu330@gatech.edu,
cassie.mitchell@bme.gatech.edu

Abstract

Lay summarization aims to simplify complex scientific information for non-expert audiences. This paper investigates the trade-off between readability and relevance in the lay summarization of long biomedical documents. We introduce a two-stage framework that attains the best readability metrics in the first subtask of BioLaySumm 2023, with 8.924 Flesch–Kincaid Grade Level and 9.188 Dale–Chall Readability Score. However, this comes at the cost of reduced relevance and factuality, emphasizing the inherent challenges of balancing readability and content preservation in lay summarization. The first stage generates summaries using a large language model, such as BART with LSG attention. The second stage uses a zero-shot sentence simplification method to improve the readability of the summaries. In the second subtask, a hybrid dataset is employed to train a model capable of generating both lay summaries and abstracts. This approach achieves the best readability score and shares the top overall rank with other leading methods. Our study underscores the importance of developing effective methods for creating accessible lay summaries while maintaining information integrity. Future work will integrate simplification and summary generation within a joint optimization framework that generates high-quality lay summaries that effectively communicate scientific content to a broader audience. Code: https://github.com/iah3/readability-summarization

1 Introduction

The burgeoning volume of biomedical literature in recent years has posed significant challenges for researchers, healthcare professionals, and the general public in staying abreast of the wealth of information generated. The task of manually summarizing long-form documents has become increasingly impractical, requiring a disproportionate amount of effort and domain-specific knowledge (Alomari et al., 2022; Adams et al., 2023; Phang et al., 2022; Al-Hussaini et al., 2022; Zhang et al., 2022a). Automatic text summarization, which seeks to distill source texts while retaining their core ideas, continues to be a demanding task, especially with long, content-rich documents laden with domain-specific complexities (Guo et al., 2022; Cao and Wang, 2022; Zhang et al., 2022b; Mao et al., 2022; Manakul and Gales, 2021).

As such, there is an urgent need to develop effective summarization techniques tailored for extensive biomedical documents to cater to diverse audiences (Goldsack et al., 2022; Ondov et al., 2022; Moro et al., 2022; Bishop et al., 2022). Meanwhile, computational complexity persists as a major obstacle, with specialized hardware potentially paving the way for more energy-efficient implementations (Gong et al., 2022; Hah et al., 2022; Athena et al., 2022a,b; West et al., 2023). As we continue to tackle these challenges, the field is poised for advancements that will fundamentally reshape how we interact with and benefit from the biomedical literature.

Lay summarization simplifies and distills complex scientific information into an accessible format for non-experts (Goldsack et al., 2023, 2022). It is vital for bridging the gap between specialized knowledge and the broader community. Controlled summarization can further enhance the accessibility of biomedical research findings by ensuring that generated summaries are both informative and comprehensible through readability. Readability-controlled summarization can maximize the use of scientific knowledge and allow various stakeholders to make informed decisions in healthcare and research (Luo et al., 2022). Generating lay summaries for long documents poses unique challenges due to the inherent complexity of the subject matter and the specialized language used in the original documents. Balancing the simplification of language with the preservation of accurate and

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 592–601
July 13, 2023 ©2023 Association for Computational Linguistics

relevant information is crucial. Reducing jargon and technical terminology may lead to the loss of essential details or the introduction of errors.

The BioNLP 2023 Workshop at ACL recently introduced a new shared task BioLaySumm, focusing on lay summarization of biomedical research articles (Goldsack et al., 2023). It comprises two subtasks with distinct objectives. The first subtask aims to generate lay summaries from PLOS and eLife articles, including the abstract (Goldsack et al., 2022), striving to maximize relevance and factuality metrics while minimizing readability score metrics. The second subtask focuses on readability-controlled summarization, which seeks to maximize relevance and factuality scores and generate both abstracts and lay summaries with readability levels comparable to the target summaries.

This paper investigates techniques for generating lay summaries and readability-controlled abstracts for long biomedical documents. We focus on their effectiveness in maintaining relevance, factuality, and readability. The ultimate goal is to facilitate knowledge dissemination and empower diverse audiences to engage with complex scientific information (Goldsack et al., 2023). A multi-step approach involving Bidirectional and Auto-Regressive Transformers (BART) (Lewis et al., 2020) and Multilingual Unsupervised Sentence Simplification (MUSS) (Martin et al., 2022) obtains the highest readability scores in the first subtask of generating lay summaries at the cost of lower relevance and factuality. In the second subtask, an approach based on Local, Sparse and Global (LSG) attention (Condevaux and Harispe, 2022) obtains the highest readability and joint highest overall scores.

2 Related Work

In recent years, biomedical document summarization has benefitted from advancements in deep learning and language models (Zhang et al., 2019a; Wang et al., 2021). Wallace et al. (Wallace et al., 2021) investigated BART (Lewis et al., 2020) for summarization of randomized controlled trials. Sotudeh et al. (Sotudeh Gharebagh et al., 2020) improved radiology report summarization by incorporating medical ontology into a sequence-to-sequence model. Domain-specific corpora using abstracts as summaries (Cohan et al., 2018; Wang et al., 2020b), has also contributed to the field. DeYoung et al. (DeYoung et al., 2021) examined the summarization of systematic reviews based on cited clinical trials. In contrast, Guo et al. (Guo et al., 2021) combined summarization and simplification to generate plain language summaries from abstracts of systematic reviews.

2.1 Lay Summarization

Prior lay summarization work primarily originates from the Shared Tasks at Scholarly Document Processing 2020: LaySumm (Chandrasekaran et al., 2020). AUTH (Gidiotis et al., 2020) employs a PEGASUS-based method to compress and rewrite article abstracts, fine-tuning the model to generate lay summaries. Dimsum (Yu et al., 2020) generates summaries using a joint extractive and abstractive approach, leveraging BART encoder and training with both extractive and abstractive summarization objectives. Kim (Kim, 2020) primarily utilizes the PEGASUS model, combining it with a BERT-based extractive model and incorporating readability metrics to enhance summary quality. Reddy et al. (Reddy et al., 2020) adopt an unsupervised extractive sentence classification method using variants of the maximum marginal relevance metric. Summaformers (Ghosh Roy et al., 2020) leverages the BART model, trained on the CNN/Dailymail dataset and fine-tuned on the LaySumm corpus. Mishra et al. (Mishra et al., 2020) employs a standard encoder-decoder framework for abstractive summarization based on BERT fine-tuned on the CNN/Dailymail dataset. Chaturvedi et al. (Chaturvedi et al., 2020) uses a two-stage pipeline involving extractive summarization, sentence extraction, and BART model-based summarization of selected text segments. However, these works were only evaluated on ROUGE score (Chaganty et al., 2018; Kryscinski et al., 2019). The recent study by Goldsack et al. (Goldsack et al., 2022) revealed that employing extractive methods or merely utilizing the abstract can lead to elevated ROUGE scores while sacrificing readability. Furthermore, the research demonstrated the capacity of a BART-based model (Lewis et al., 2020) to generate lay summaries with both high relevance scores and low readability scores, thus achieving the desired outcome.

2.2 Readability, Relevance, and Factuality

Readability, which reflects the ease of understanding a text, is influenced by factors like lexical and syntactic complexity, discourse cohesion, and background knowledge (Crossley et al., 2017). In this study, we evaluate readability using Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975; Flesch, 2007) previously employed in lay summarization research (Guo et al., 2021), and the Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948; Chall and Dale, 1995).

Relevance metrics like ROUGE (Chaganty et al., 2018; Kryscinski et al., 2019) and BERTScore (Zhang et al., 2019b) assess if a summary captures the source's main ideas. Factual consistency metrics evaluate summary-source consistency (Goyal and Durrett, 2020; Wang et al., 2020a). Despite high factual error rates in short-document model-generated summaries (Cao et al., 2018; Maynez et al., 2020), efforts have focused on developing effective factuality metrics (Honovich et al., 2021; Xie et al., 2021; Ribeiro et al., 2022). ROUGE scores best correlate with human relevance scores, providing a platform for benchmarking long document abstractive models (Koh et al., 2022). According to Koh et al., the best overall correlation with human factual consistency scores is achieved by fine-tuned BARTScore (Koh et al., 2022).

3 Data

The shared task leverages data from two principal sources, the Public Library of Science (PLOS) (Goldsack et al., 2022, 2023; Luo et al., 2022) and eLife (Goldsack et al., 2022, 2023). Each of these datasets comprises biomedical research articles, paired with their technical abstracts and lay summaries written by experts. As discussed in the preceding section, the utility of each type of summary differs depending on the subtask. Moreover, the lay summaries in each dataset exhibit several distinct characteristics; for more comprehensive details, readers are referred to (Goldsack et al., 2022, 2023; Luo et al., 2022).

The PLOS dataset, the larger of the two, contains 24,773 instances designated for training and 1,376 instances for validation, applicable to both subtasks. In contrast, the eLife dataset consists of 4,346 training instances and 241 validation instances.

The test data for subtask 1 includes 142 articles each from PLOS and eLife. Meanwhile, the test data for subtask 2 comprises 142 PLOS articles, distinct from those used in subtask 1.

It's important to note that the test sets allocated for both subtasks are distinct from those published in any of the cited papers, and are made accessible via the CodaLab pages dedicated to the shared task. This configuration fosters a rich and diverse dataset, providing comprehensive support for the objectives of the shared task.

3.1 Metrics

In order to effectively gauge the performance of the proposed models, a range of comprehensive metrics were assigned in the evaluation.

To measure the relevance of the summaries to their source documents, ROUGE-1, ROUGE-2, and ROUGE-L metrics (Chaganty et al., 2018; Kryscinski et al., 2019) were assigned. These metrics provide insights into summary quality by comparing them with human-authored references, specifically by calculating the overlap of n-grams between the generated summary and the source text. ROUGE-1, ROUGE-2, and ROUGE-L evaluate the overlap of unigrams, bigrams, and longest common sequences, respectively.

BERTScore (Zhang et al., 2019b) was also assigned for relevance evaluation. BERTScore uses BERT embeddings to calculate semantic similarity between generated summaries and source texts, providing a more refined evaluation than raw n-gram overlap.

To evaluate readability, Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975; Flesch, 2007) and Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948; Chall and Dale, 1995) were used. These metrics measure text understanding ease, considering elements such as syntactic complexity, lexical diversity, and sentence length. FKGL estimates the US school grade level required to understand the text, while DCRS measures complexity based on word familiarity and sentence length.

The factual consistency of the produced summaries is an essential aspect in summarization tasks. For this purpose, BARTScore (Koh et al., 2022) was employed. This metric measures the consistency between summary and source, pinpointing factual inconsistencies and ensuring that the summaries accurately reflect the original source content. As per the findings of Koh et al., BARTScore has the best overall correlation with human factual consistency scores, making it a reliable choice for this evaluation.

The application of these metrics in the evaluation process facilitates a comprehensive analysis of the generated summaries' quality, relevance, read-

Table 1: Lay summarization scores on the test set on CodaLab. The proposed methods (PD 1, PD2, PD3) are delineated from the baseline and the top-ranked team, MDC, by a line. Lower scores are favorable for FKGL and DCRS, while higher scores are advantageous for the other metrics, as illustrated by the arrows.

	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	BERTScore ↑	FKGL ↓	DCRS ↓	BARTScore ↑
Baseline	0.470	0.145	0.437	0.864	12.069	10.249	-0.831
MDC	0.482	0.155	0.449	0.871	12.937	10.206	-1.177
PD 1 - BART[1]	0.494	0.159	0.460	0.859	13.096	10.122	-2.331
PD 2 - BART[1] + MUSS	0.475	0.135	0.448	0.854	8.924	9.188	-3.230
PD 3 - LSG[2]	0.473	0.148	0.438	0.857	12.488	9.986	-2.178

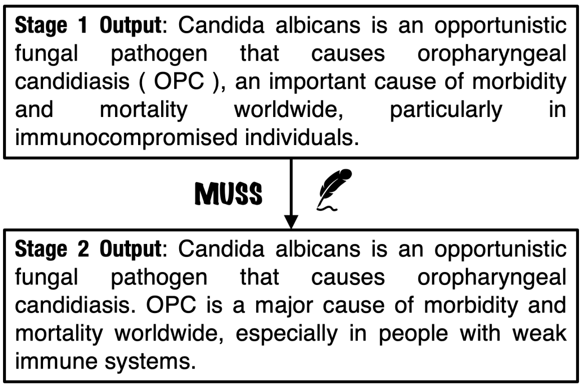

Figure 1: Sentence Simplification in Lay Summary Generation for Subtask 1 using MUSS (Martin et al., 2022). Although readability improves, relevance and factuality suffers due to the inclusion of an abbreviation without clarification.

Table 2: Final rankings of the proposed methods for lay summarization. PD 1, PD 2, and PD 3 denote the three methods proposed. PD 2 obtains the best overall readability score.

	Relevance	Readability	Factuality	Overall
PD 1-BART[1]	4	11	14	9
PD 2-BART[1]+MUSS	13	1	19	11
PD 3-LSG[2]	12	8	11	10

ability, and factual accuracy.

4 Lay Summarization

The generation of lay summaries was executed through a two-stage method. Prior to delving into this method, it's important to outline the baseline approach from which it sprang. This involved using a base BART model (Lewis et al., 2020), pretrained on the CNN Daily Mail (Hermann et al., 2015; See et al., 2017), with an input and output token length of 1024[1]. Any articles exceeding 1024 tokens in length were truncated. This approach was the most successful of the three proposed methods when generating abstractive lay summaries, achieving the highest overall scores, primarily due to its superior relevance score, which is consistent with the findings of Goldsack et al. (Goldsack et al., 2022). This approach is referred to as PD 1 (Pathology Dynamics) in Tables 1 and 2.

However, this BART-based approach fell short in terms of readability scores. To remedy this, the two-stage method was implemented. In the first stage, a large language model such as BART (Lewis et al., 2020) or LSG (Condevaux and Harispe, 2022) was trained to generate lay summaries based on the articles, which incorporated abstracts, containing the most important information in a concise format. This eliminated the need for longer input token lengths required in Subtask 2, which did not include the abstract in the input.

In the second stage, the two-stage approach used a zero-shot sentence simplification method called MUSS (Martin et al., 2022) to enhance the readability of the generated lay summaries. This approach, referred to as PD 2 in Tables 1 and 2, achieved the best readability score among all the participating teams in the subtask, attesting to the effectiveness of integrating MUSS (Martin et al., 2022) into the process.

Despite this success, Figure 1 indicates certain limitations with MUSS (Martin et al., 2022). While readability was improved, certain aspects of clarity were compromised, such as the detachment of the abbreviation OPC from its full form, oropharyngeal candidiasis. As a result, the summary's relevance and factuality scores dipped below that of the base model, emphasizing the need for a careful balance between simplification techniques and the maintenance of essential information.

The final method involved the use of an LSG attention model (Condevaux and Harispe, 2022) with an increased input token length of 4096 for the encoder to generate lay summaries, while keeping

[1]facebook/bart-large-cnn
[2]ccdv/lsg-bart-base-4096
[3] mrm8488/t5-base-finetuned-summarize-news

Table 3: Readability-controlled summarization scores on the test set on CodaLab. The proposed method is distinguished from other approaches by a line. Lower scores are preferable for FKGL and DCRS, while higher scores are desired for the remaining metrics, as denoted by the arrows.

Team	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	BERTScore ↑	FKGL ↓	DCRS ↓	BARTScore ↑
Baseline	0.409	0.116	0.369	0.855	2.396	0.931	-0.978
NCUEE-NLP	0.451	0.140	0.412	0.855	2.048	0.934	-2.110
LHS712EE	0.442	0.130	0.405	0.855	2.263	0.936	-1.140
Pathology Dynamics	0.451	0.138	0.410	0.853	2.107	0.823	-1.568

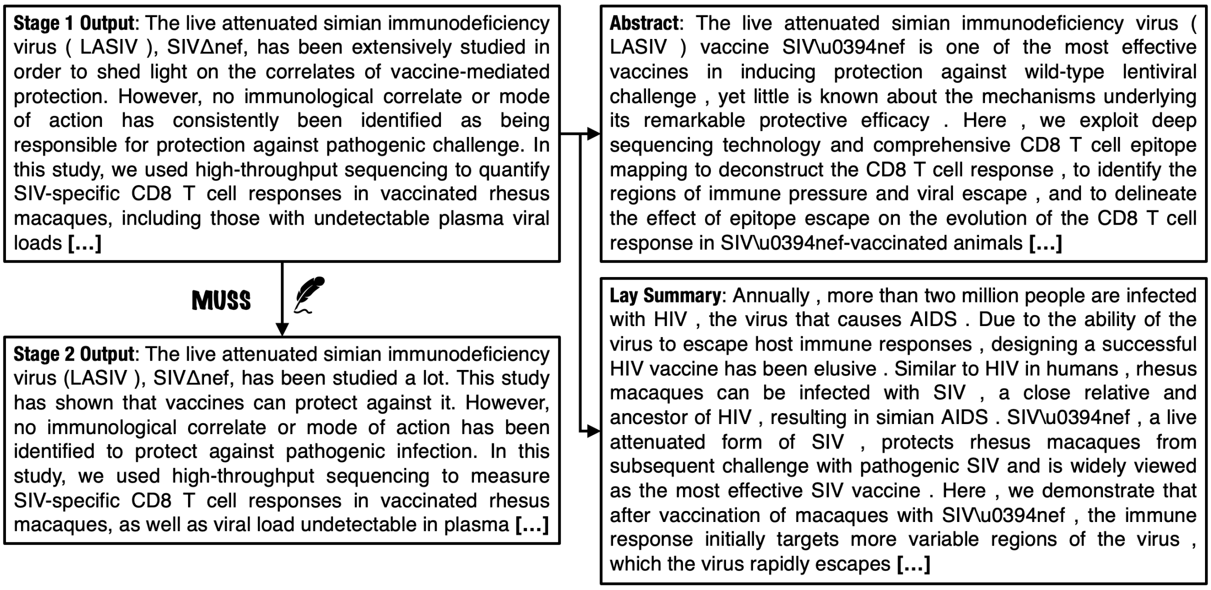

Figure 2: Generation of abstracts and lay summaries for Subtask 2. The figures on the left depict the results from the two-stage approach. Text boxes on the right display the original abstract of the article and the desired lay summary. Upon manual review and metric evaluation using the validation set, it is noted that the output from stage 2 diverges from the target lay summary. As indicated by the arrows, the output from stage 1 therefore acts as both the lay summary and the abstract for the purposes of evaluation.

the decoder's output token length at 1024[2]. This approach, denoted as PD 3 in Table 1, yielded the most balanced results and achieved the highest factuality among the proposed methods, as seen in Table 2. Thanks to its efficient attention mechanism, the LSG model's training was as swift as BART's, despite a longer input token length.

5 Readability Controlled Summarization

This subtask was designed to yield lay summaries and abstracts from other sections of the article using a singular model. In order to address this intricate task, we initiated our methodology with a pre-processing phase designed to create a synthetic dataset. This was achieved by duplicating each article and correlating it with the corresponding abstract and lay summary as outputs. As a result, this tactic effectively doubled the number of articles available for training in our dataset, representing a significant enhancement over the volume offered by the original dataset.

Following the augmentation of our dataset, we utilized this expanded synthetic dataset to train a model based on LSG (Condevaux and Harispe, 2022). This particular model was adept at generating a summary that seamlessly integrated components from both the lay summary and the article. One notable characteristic of this subtask was the absence of an abstract in the input articles. This invariably meant that crucial information was scattered throughout the entirety of the article, thereby calling for a greater input token length.

To account for this, we engaged an LSG-based model (Condevaux and Harispe, 2022), specifically configured with an encoder that accommodated 4096 input tokens and a decoder that handled 1024 output tokens[2]. This specialized setup surpassed the performance metrics exhibited by T5[3] (Raffel

Table 4: Final ranking of readability-controlled summarization. Pathology Dynamics refer to our proposed method using LSG attention.

Team	Relevance	Readability	Factuality	Overall Rank
Baseline	4	4	1	4
NCUEE-NLP	1	2	4	1
LHS712EE	2	3	2	1
Pathology Dynamics	3	1	3	1

et al., 2020) and BART[1] (Lewis et al., 2020) models with 1024 input tokens , as per the results from the validation set. Consequently, this led to the selection of the LSG-based model for the following stages of the task, underlining its capability to deliver effective summarization for long biomedical documents. The results are presented in Tables 3 and 4.

Moving forward, we incorporated the two-stage modeling approach, as described in the previous task, into this current subtask. In the initial stage, the LSG-based model (Condevaux and Harispe, 2022) was tasked with generating a summary intended to act as the abstract. The subsequent stage hinged on the use of MUSS (Martin et al., 2022) to produce the lay summary. However, an in-depth inspection of the generated summaries coupled with an analysis of the validation set scores led us to the realization that the output from the first stage of the model served as a more effective lay summary than that produced by MUSS. We found that the MUSS output was excessively simplified when juxtaposed against the target lay summary.

Due to this observation, we opted for the utilization of the output from the first stage to fulfill the dual roles of both the abstract and the lay summary. The objective of this task was to achieve readability scores that more closely aligned with the target values, as opposed to pursuing the lowest possible scores. Given this context, the application of MUSS resulted in a decline in readability scores for this subtask.

Figure 2 presents the opening sentences of the outputs from the first and second stages for a representative article. This visual comparison highlights the propensity of MUSS (Martin et al., 2022) to oversimplify the summary, causing it to diverge from the intended lay summary. As a result, the utilization of the same output from the first stage for both the abstract and lay summary culminated in superior scores, even in terms of readability. This is because the readability score for this subtask involved a comparison between the generated lay summary and the original version. This points to the inherent challenge of striking a balance between readability and the retention of crucial content when employing simplification strategies for the generation of lay summaries.

6 Conclusion and Future Work

In conclusion, this paper explored the intricate trade-offs between readability, relevance, and factuality in lay summarization. We have highlighted the inherent challenges associated with transforming complex scientific information into an accessible format for non-expert audiences. Our proposed two-stage framework attains state-of-the-art readability metrics; however, this is achieved at the cost of reduced relevance and factuality. These findings emphasize the necessity of striking a balance between readability and content preservation when creating lay summaries.

Future work will focus on integrating simplification and summary generation within a joint optimization framework. This approach aims to overcome the trade-offs identified in our study and enable the generation of high-quality lay summaries without sacrificing readability, relevance, or factuality. More effective lay summarization methods can be developed by considering both simplification and summarization as complementary processes.

Code Availability

Code is available on GitHub: `https://github.com/iah3/readability-summarization`

Limitations

This study is constrained by limited input token length due to hardware memory limitations and lengthy training times. Even with the LSG attention mechanism's efficiency, this inadequacy persists in both subtasks. Longer token length could improve summary relevance and factuality. Particularly in the second subtask, where the abstract is absent, this poses a challenge in generating a summary from sections with high information density.

Acknowledgements

This research was funded by the NSF CAREER grant 1944247 to C.M, the NIH grant U19-AG056169 sub-award to C.M., the Morningside Center for Innovative and Affordable Medicine at Emory University via the Brown Innovation to Market Fund to C.M.

References

Griffin Adams, Bichlien H Nguyen, Jake Smith, Yingce Xia, Shufang Xie, Anna Ostropolets, Budhaditya Deb, Yuan-Jyue Chen, Tristan Naumann, and Noémie Elhadad. 2023. What are the desired characteristics of calibration sets? identifying correlates on long form scientific summarization. *arXiv preprint arXiv:2305.07615*.

Irfan Al-Hussaini, Davi Nakajima An, Albert J. Lee, Sarah Bi, and Cassie S. Mitchell. 2022. Ccs explorer: Relevance prediction, extractive summarization, and named entity recognition from clinical cohort studies. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5173–5181.

Ayham Alomari, Norisma Idris, Aznul Qalid Md Sabri, and Izzat Alsmadi. 2022. Deep reinforcement and transfer learning for abstractive text summarization: A review. *Computer Speech & Language*, 71:101276.

Fabia F Athena, Matthew P West, Pradip Basnet, Jinho Hah, Qi Jiang, Wei-Cheng Lee, and Eric M Vogel. 2022a. Impact of titanium doping and pulsing conditions on the analog temporal response of hafnium oxide based memristor synapses. *Journal of Applied Physics*, 131(20):204901.

Fabia F Athena, Matthew P West, Jinho Hah, Riley Hanus, Samuel Graham, and Eric M Vogel. 2022b. Towards a better understanding of the forming and resistive switching behavior of ti-doped hfo x rram. *Journal of Materials Chemistry C*, 10(15):5896–5904.

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

Shuyang Cao and Lu Wang. 2022. HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807, Dublin, Ireland. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224, Online. Association for Computational Linguistics.

Rochana Chaturvedi, Saachi ., Jaspreet Singh Dhani, Anurag Joshi, Ankush Khanna, Neha Tomar, Swagata Duari, Alka Khurana, and Vasudha Bhatnagar. 2020. Divide and conquer: From complexity to simplicity for lay summarization. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 344–355, Online. Association for Computational Linguistics.

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

Charles Condevaux and Sébastien Harispe. 2022. Lsg attention: Extrapolation of pretrained transformers to long sequences.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

Sayar Ghosh Roy, Nikhil Pinnaparaju, Risubh Jain, Manish Gupta, and Vasudeva Varma. 2020. Summaformers @ LaySumm 20, LongSumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 336–343, Online. Association for Computational Linguistics.

Alexios Gidiotis, Stefanos Stefanidis, and Grigorios Tsoumakas. 2020. AUTH @ CLSciSumm 20, LaySumm 20, LongSumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 251–260, Online. Association for Computational Linguistics.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

N. Gong, M.J. Rasch, S.-C. Seo, A. Gasasira, P. Solomon, V. Bragaglia, S. Consiglio, H. Higuchi, C. Park, K. Brew, P. Jamison, C. Catano, I. Saraf, F.F. Athena, C. Silvestre, X. Liu, B. Khan, N. Jain, S. Mcdermott, R. Johnson, I. Estrada-Raygoza, J. Li, T. Gokmen, N. Li, R. Pujari, F. Carta, H. Miyazoe, M.M. Frank, D. Koty, Q. Yang, R. Clark, K. Tapily, C. Wajda, A. Mosden, J. Shearer, A. Metz, S. Teehan, N. Saulnier, B. J. Offrein, T. Tsunomura, G. Leusink, V. Narayanan, and T. Ando. 2022. Deep learning acceleration in 14nm cmos compatible reram array: device, material and algorithm co-optimization. In *2022 International Electron Devices Meeting (IEDM)*, pages 33.7.1–33.7.4.

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, Online. Association for Computational Linguistics.

Mandy Guo, Joshua Ainslie, David C 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.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Jinho Hah, Matthew P West, Fabia F Athena, Riley Hanus, Eric M Vogel, and Samuel Graham. 2022. Impact of oxygen concentration at the hfo x/ti interface on the behavior of hfo x filamentary memristors. *Journal of Materials Science*, 57(20):9299–9311.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seungwon Kim. 2020. Using pre-trained transformer for better lay summarization. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 328–335, Online. Association for Computational Linguistics.

J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated readability index. *Naval Technical Training Command Millington TN Research Branch*.

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022. How far are we from robust long abstractive summarization? *arXiv preprint arXiv:2210.16732*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

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

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

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

Santosh Kumar Mishra, Harshavardhan Kundarapu, Naveen Saini, Sriparna Saha, and Pushpak Bhattacharyya. 2020. IITP-AI-NLP-ML@ CL-SciSumm 2020, CL-LaySumm 2020, LongSumm 2020. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 270–276, Online. Association for Computational Linguistics.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. 2022. Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 180–189.

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

Jason Phang, Yao Zhao, and Peter J Liu. 2022. Investigating efficiently extending transformers for long input summarization. *arXiv preprint arXiv:2208.04347*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Saichethan Reddy, Naveen Saini, Sriparna Saha, and Pushpak Bhattacharyya. 2020. IIITBH-IITP@CL-SciSumm20, CL-LaySumm20, LongSumm20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 242–250, Online. Association for Computational Linguistics.

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, Seattle, United States. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905, Online. Association for Computational Linguistics.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605.

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

Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, et al. 2021. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020b. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Matthew P West, Fabia F Athena, Samuel Graham, and Eric M Vogel. 2023. Bias history impacts the analog resistance change of hfox-based neuromorphic synapses. *Applied Physics Letters*, 122(6):063502.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiezheng Yu, Dan Su, Wenliang Dai, and Pascale Fung. 2020. Dimsum @LaySumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 303–309, Online. Association for Computational Linguistics.

Haopeng Zhang, Semih Yavuz, Wojciech Kryściński, Kazuma Hashimoto, and Yingbo Zhou. 2022a. Improving the faithfulness of abstractive summarization via entity coverage control. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 528–535.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019a. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 789–797, Hong Kong, China. Association for Computational Linguistics.

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

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022b. Summn: A multi-stage summarization framework for long input dialogues and documents: 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.