Biomedical Aspect-based Summarization with Personalized LLMs

Joe D. Menke
School of Information Sciences
University of Illinois at Urbana-Champaign
jmenke2@illinois.edu

Rahul Chappidi Venkata
School of Information Sciences
University of Illinois at Urbana-Champaign
rc46@illinois.edu

***Abstract*—Large language models (LLMs) can be used to summarize lengthy articles into a more concise form, allowing more efficient knowledge transfer. This is particularly valuable in fields like biomedicine, where millions of articles are published annually and yet clinicians and researchers are expected to keep up to date. Personalized summarization adds another layer by tailoring summaries to individual preferences or specific professional needs, which may further enhance knowledge transfer. In this study, we create a dataset for aspect-based summarization using four different author-generated summaries extracted from biomedical review articles. Additionally, we evaluate various models using a suite of automated metrics. Based primarily on the SummaC factuality metric, the persona-based system without reflection performed best on the abstract (SummaC = 0.61) and plain language summarization (SummaC = 0.64), while, the baseline surprisingly appears to have performed the best for the implications for research (SummaC = 0.75) and for practice summaries (SummaC = 0.68). Future work aims to expand our experiments and improve the quality of our evaluation through additional metrics that incorporate the goal of each task as well as through manual evaluation.**

***Index Terms*—biomedicine, large language models, aspect-based summarization, personalization**

I. Introduction

Within natural language processing (NLP), the task of generating a concise version of documents while still preserving key information is a common task, known as summarization. One of most common variants of this task is known as Plain Language Summarization (PLS), which attempts to simplify text during the summarization process. PLS essentially summarizes text personalized to someone without some particular knowledge. While important on its own, PLS highlights the fact that a single summary may not work for every person as different users will need different levels of simplification or focus depending on their background or goal. For example, a biomedical researcher compiling a systematic review for randomized controlled trials may pay significantly more attention to aspects of the methods section relating to the rigor and reproducibility of the study (i.e., any poor study designs leading to biased results), whereas a clinician reading the same articles may focus more on the results and implications. This thinking aligns well with aspect-based summarization, a subtask of summarization, which aims to generate multiple summaries for a single document each considering different user perspectives. Aspect-based summarization for biomedical research articles has not been deeply explored to our knowledge.

The purpose of this work is to bridge this gap by incorporating user-specific perspectives into the summarization process of biomedical research articles. Firstly, we create a dataset that features summaries from four different perspectives for biomedical review articles. Using a suite of evaluation metrics, we develop a multi-agent summarization pipeline using large language models (LLMs) in efforts to generate better outputs with respect to different user groups and their various needs. Ultimately, the goal of this work is to generate a more robust framework for summarization as well as a benchmark dataset on which future personalized summarization systems may be evaluated for comparison within the biomedical domain.

II. Related Works

Prior work on summarization methods extensively explores both extractive and abstractive approaches, with growing attention toward hybrid models. Extractive summarization involves selecting the most important sentences from the source text based on statistical features, such as term frequency or sentence centrality, which rank sentences for inclusion in the summary. These methods, while computationally efficient and capable of preserving the factuality of the original content, often result in summaries that lack semantic cohesion and narrative flow, as they simply concatenate extracted sentences without generating new text. Graph-based methods such as TextRank [1] or LexRank [2], often applied in extractive summarization, employ sentence-similarity graphs but struggle with semantic equivalency and the "dangling anaphora" problem, which reduces coherence. Overall, these methods may generate summaries that maintain high factuality with reference to their source text, however, they are restrictive in that they cannot be adapted based on the needs of the user, e.g., in the case of PLS. Additionally, they are not able to synthesize information, which may be needed to best summarize a biomedical research article, as they simply extract existing sentences.

In contrast, abstractive summarization aims to generate summaries that are more coherent and resemble human-generated texts. With the rise of deep learning, encoder-decoder architectures such as transformer-based [3] models like BART [4] and T5 [5] have enhanced the performance of abstractive methods

by enabling better contextual understanding of the source text [6]. However, these generative models do present their own technical challenges, particularly in maintaining factual accuracy. For example, large language models sometimes introduce knowledge not referenced within the source text at all, usually referred to as hallucinations [7], [8]. They may also oversimplify or subtly alter the meaning of text, which can result in a loss of information. As such, it is important to evaluate these models not just on metrics related to content similarity and style, but also on factuality [9], [10].

More recently, research has started to apply these methods within domain-specific variations of summarization, like in the biomedical field. For example, Cohan et al. created a summarization dataset where the reference summaries are research abstracts and the source text is the full-text of the article [11]. As previously mentioned, a common variant of biomedical summarization is PLS. One of the goals for PLS is to generate lay summaries that simplify complex biomedical research for non-expert readers, preserving technical accuracy while enhancing accessibility. Abstractive summarization models can be designed to produce summaries in plain language, often focusing on readability metrics such as Flesch-Kincaid Grade Level [12] and Dale-Chall Readability Score [13] to ensure that the output is suitable for non-specialists. However, the best metrics for PLS (and other summarization tasks) are still a subject of debate with the most recent work recommending a suite of metrics [14]. Tasks such as BioLaySumm [15] emphasize the need for readability-controlled summarization, where models are trained to generate both technical and lay summaries from the same source text. These models usually rely on fine-tuned transformer architectures like Longformer Encoder-Decoder (LED) [16] and BioBART [17], with modifications to ensure the balance between factual accuracy and ease of comprehension, although LLM methods were also explored. This task demonstrates the importance of tailoring models to meet audience-specific needs in order to address the broader challenge of making information more accessible to others.

General summarization and PLS are the primary summarization tasks being explored within biomedicine right now. However, biomedicine is read and used by individuals with far more nuanced use cases. Limiting summarization to these two tasks is limiting the utility of abstractive summarization within this rich and diverse source of information. Researchers come from a range of different fields and subfields that may or may not be related to information in articles. Reducing their information needs down to either expert or layperson is overly simplistic. Guo et al. has recently started to explore this from the perspective of summarizing text based on the knowledge of individual researchers [18]. However, this personalization largely works to mitigate the negative effects of researchers reading outside of their specialized field, rather than personalizing based on different tasks.

Aspect-based summarization was initially proposed to better summarize product reviews [19], [20]. However, its overall goal of summarizing text with consideration of a user's view aligns well with summarizing biomedical research articles. Recent work has been done to develop large-scale, aspect-based summarization datasets using wikipedia [21], [22]. Most recently, ACLSum, a multi-aspect summarization dataset focusing on scientific papers, was introduced [23]. This dataset was developed using research articles published in ACL venues and features 3 different summary types: challenges, approaches, and outcomes. While valuable, this dataset largely focuses on natural language processing, which can be quite different than biomedicine (both in terms of content and reader perspectives).

Here, this work attempts to focus on biomedical summarization, specifically, by developing an aspect-based summarization task that starts to consider the underlying reasons a researcher may be reading a biomedical review article, for example, a researcher within a different field expanding their knowledge, or a clinician simply trying to stay up to date on the best practices or treatments for a particular disease.

III. Methods

The overall flow of the study is shown in Figure 1. The diagram shows three phases of this project: (1) the extraction of each summary from the full-text during the creation of this dataset, (2) generating summaries using different methods, and (3) evaluation of each generated summary. While we currently use a suite of 10 evaluation metrics, we keep the question mark to indicate that future work should still focus on how best to evaluate these generated summaries.

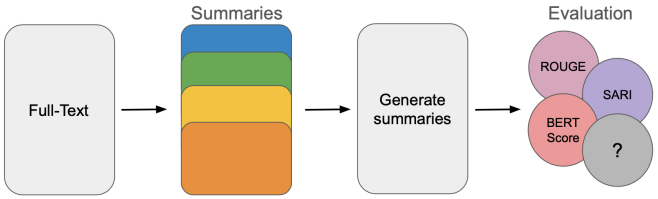

Fig. 1. Flow diagram of the overall study, including use of 4 different summary types (abstract, plain language summary, implications for research, and implications for practice) as well as some of the different evaluation metrics utilized.

A. Dataset

We identified and created our dataset, utilizing 4 article-summary pairings from different perspectives: abstract summary (a researcher from the same field interested in deep-level summary), layperson summary (a summary for a non-researcher or a researcher from a different domain), implications for research summary (a researcher or stakeholder interested in high-level conclusions), and implications for practice (a clinician interested in high-level conclusions applicable to their practice). This dataset contains 302 review articles published by the Cochrane Database of Systematic Reviews. These articles all include summaries of the 4 different perspectives included in this dataset and were downloaded through the PubMed Central Open Archives Initiative (PMC-OAI) as XML files. These XML files were then parsed to separate

the 4 summary types from the other article text. The extracted summaries were removed from the source text used to initially generate summaries as we did not want the model to simply extract text but rather synthesize this information. This dataset is split roughly into training (60%, n=194), validation (30%, n=84), test (10%, n=25) splits for development of our system. Test was split by year; any article published in 2024 was placed into the test set as instances without any sort of data leakage for a majority of evaluated LLM models (e.g., GPT-4 models have a knowledge cut-off of October 2023). The remaining articles were split randomly into training and validation sets.

To measure the lengths of the summaries as well as the rest of text (i.e., full-text with summaries removed), we utilized GPT-4o's tokenizer. GPT-4o's tokenizer was used as it is readily available and is expected to be similar to tokenizers used by many popular LLMs.

B. Models

The experimental framework for our model is shown in Figure 2. The methods we evaluate all use task-specific prompts to generate summaries, which are then compared against task-specific reference summaries for evaluation.

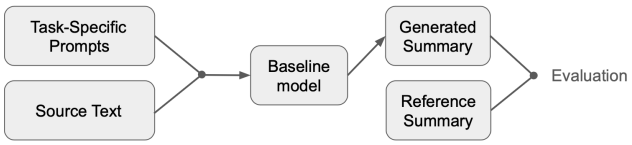

Fig. 2. The experimental framework for our experiments, which all utilize 4 different task-specific prompts (1 for each summary type). Generated summaries are then compared against task-specific reference summaries extracted previously from the full-text article.

The large language model used for testing in this work is Llama-3.1-8B-Instruct. This was chosen as it is open-access and freely available through HuggingFace. As previously alluded to in Figure **??**, the input involves task-specific prompts and input source text. These are fed into the model to generate summaries. The generated summaries are then compared to reference summaries using a suite of evaluation metrics. This pipeline allows for a structured comparison between model outputs and human-generated summaries, enabling assessment of the model's effectiveness in summarization tasks.

1) User-Persona Multi-Agent System: As currently implemented, the user-persona based multi-agent summarization system features 4 agents: a persona-generation agent, a persona-summarizer, a persona-critic, and a persona-editor. An overview is shown in Figure 3. Current experiments represents this framework at its most basic. Future work may extend this to incorporate multiple reflection steps (i.e., iterating multiple times between the persona-critic and the persona-editor) as well as tooling (allow editors access to the original source article). The input into the system is the same as our baseline experiment: a description of the task as well as the textual input to be summarized. These are used to generate a persona, which is a description of someone who would read a research article with that task in mind, for example, a physician. This persona is used to initialize the summarization agent who generates the task-specific summary. The generated summary is then critiqued by the persona and then edited based on this feedback. Finally, the edited summary is outputted as the final result. As a sort of role-playing, this persona may generate better summaries, especially those meant to best serve similar users.

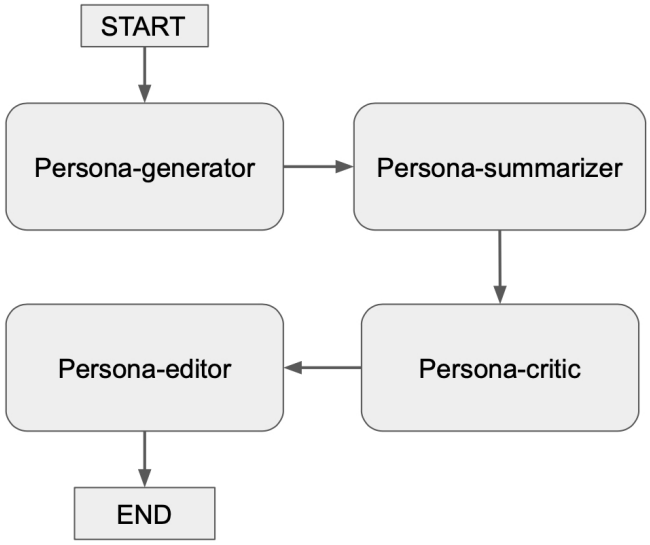

Fig. 3. An overview of our persona-based multi-agent framework. There are 4 agents: a persona-generation agent, a persona-summarizer, a persona-critic, and a persona-editor.

Overall, we experimented with 3 system architectures. The first serves as baseline and is simply the LLM with the task-specific prompt and input text. The second experiment adds the persona-generation step. This persona is then used to generate the summary. Finally, for our third experiment, we added reflection usng the critique and edit steps. In theory, future works may repeat the reflection step dynamically (multiple times or stopping based on input from another LLM).

IV. Evaluation

Based on insights from prior work [14], we evaluate our systems using a suite of metrics. For relevance, we utilize common summarization metrics including ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore [24]. These metrics measure the overlap of word pieces, words, and phrases, which evaluate how relevant each summary is (i.e., does the generated summary use similar words as the reference summary?). This is similar to previous aspect-based summarization methods, which primarily used ROUGE variants for evaluation [21], [22]. For readability, we use SARI [25], Flesch-Kincaid Grade Level [12] (FKGL), and Dale-Chall Readability Score [13] (DCRS). For SARI, a higher score indicates that an article is more readable, while for FKGL and DCRS, the score corresponds to the grade-level, so lower scores are more readable. In efforts to evaluate based on factuality, we also include SummaC [9]. We tried to utilize AlignScore [10] as well, but

found difficulty in implementing it. While in an ideal world, these factuality metrics would be most important (especially within biomedicine), we feel we must note the caveat that automated metrics regarding factuality are still not great [26], [27]. For example, Fang, Dai, and Karimi demonstrate that most common factuality metrics do not correlate well with human evaluations [26]. Similarly, Ramprasad and Wallace show that these metrics are sensitive to irrelevant text and could potentially be gamed [27]. These metrics serve as our evaluation, but future work should manually analyze generated summaries.

V. Results

A. Dataset

After removal of tables, the average token length of the rest-of-text (i.e., the full-text article with summaries removed) was 18,157.62 ± 10,469.56 tokens, demonstrating that the articles varied widely in length. The minimum token length was 4,245, while the maximum was 107,641 tokens. The first quartile was 11,098 tokens; the median was 15,614.5 tokens; finally, the third quartile was 22,297 tokens.

Regarding the summaries, the box plots for various token lengths each are shown in Figure 4. As shown, the abstract (1219.14 ± 323.33 tokens) and plain language summaries (740.35 ± 235.20 tokens) were generally larger than the other summaries - implications for practice (336.49 ± 364.07 tokens) and implications for research (394.81 ± 399.21 tokens), although these generally had longer tails. For example, implications for research had a summary with a maximum length of 3686 tokens. Intuitively, it makes sense that abstracts and PLS have less outliers as they are generally highly structured with specific instructions relating to length whereas other summaries, especially implications-related ones (which are entirely unrelated to the abstract) may be less defined. The plain language summary is usually based on the abstract, so it makes sense that it would be more uniform in length, similar to the abstract. Additionally, the PLS's shorter length in comparison to the abstract makes sense as it represents a simplified version of the abstract.

B. Models

The results from the baseline single-prompt summarization generation using the Llama3.1-8B-Instruct model show varied performance across different categories. Other models were not tested due to various challenges outlined in the discussion section of this paper. The model achieved relatively low BLEU scores across all tasks, with the highest being 0.122 in Practical Implications, indicating limited overlap with reference summaries in terms of exact phrasing (i.e., n-grams). However, the model performed significantly better on the SARI metric, averaging a score of 42.04, which suggests a moderate ability to generate simplified or adjusted text that aligns with the desired summary requirements. The BERTScore-F1, which measures semantic similarity, averaged 0.8575, showing strong alignment with reference meanings despite low surface-level overlap. These results highlight the model's effectiveness in producing semantically similar summaries, although it struggles with generating exact matches to reference texts. This performance may suggest that while the model captures essential ideas well, improvements may still be needed in order to generate more precise, structured outputs. More comparisons are needed though (e.g., summaries generated by randomly selecting sentences from the rest of text) to better contextualize these results.

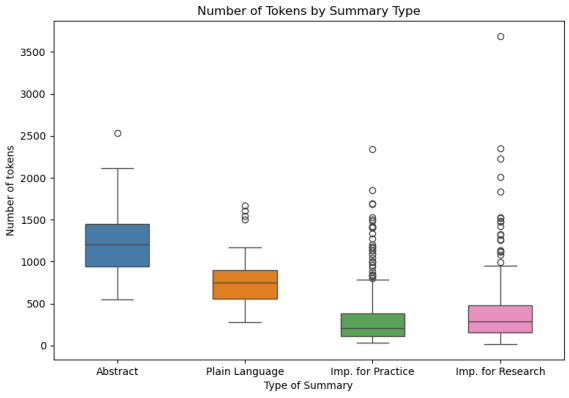

Fig. 4. Boxplots showing the various token lengths for each distinct summary type. Blue represents the abstract. Orange represents the plain language summary. Green represents the implications for practice summary. Pink represents the implication for research summary.

VI. Discussion

The creation of this dataset allows others to start developing systems for biomedical aspect-based summarization. To our knowledge, this dataset is the first of its kind, featuring multiple aspects. Previous work focused on abstract [11] or plain language summary generation [15], [28]. This dataset alsoo opens the door to other aspects in biomedicine as well. For example, a summary based on the rigor and reproducibility of a study (e.g., does the study design bias the results in someway?) would be helpful as it would ground the credibility of a research article quickly for readers.

Regarding the LLM system(s) we used for summarization, more work needs to be done. While we obtained initial results on our test set, the experiments use a relatively small LLM with only 8 billion parameters. Future work should expand to analyze even larger models (e.g., 70B) and models from other "families" (e.g., GPT-4, Gemini, etc.) to measure their impact. Additionally, we could also alter numerous components within the summarization system to potentially enhance performance. The suite of metrics helps to raise the point that automated evaluation metrics need to be better, especially if we want to use this sort of summarization-based framework within biomedicine. This is the case because it is not entirely clear which experiment performed best. It depends on which metric(s) are being used for evaluation. Manual evaluation is obviously useful, but it is also slow and costly. To be of

TABLE I

PERFORMANCES ON THE TEST SET (N=25) ACROSS DIFFERENT SUMMARY TYPES AND EXPERIMENTS USING THE LLAMA3.1-8B-INSTRUCT MODEL. BASE REFERS TO THE BASELINE EXPERIMENT WHERE AN LLM WAS ASKED TO GENERATE A TASK-SPECIFIC SUMMARY USING A MANUALLY DEVELOPED PROMPT. PERSONA REFERS TO THE EXPERIMENT USING A PERSONALIZED LLM TO GENERATE THE SUMMARY. REFLECTION REFERS TO THE EXPERIMENT WHERE A PERSONALIZED LLM GENERATES A SUMMARY, A CRITIQUE, AND A REVISED SUMMARY.

Experiment	Metric	Abstract	Plain Language Summary	Implications for Research	Implications for Practice
Base	ROUGE-1 ↑	0.03	0.05	0.04	0.04
	ROUGE-2 ↑	0.01	0.01	0.01	0.009
	ROUGE-L ↑	0.03	0.04	0.03	0.03
	BERTScore-P ↑	0.87	0.89	0.88	0.85
	BERTScore-R ↑	0.83	0.87	0.85	0.91
	BERTScore-F1 ↑	0.84	0.90	0.86	0.87
	SummaC ↑	0.56	0.61	0.75	0.68
	SARI ↑	40.6	40.0	37.5	37.6
	Flesch-Kincaid Grade Level (FKGL) ↓	14.4	11.5	14.2	14.3
	Dale-Chall Readability Score (DCRS) ↓	8.5	8.98	8.97	9.17
+ Persona	ROUGE-1 ↑	0.03	0.04	0.04	0.03
	ROUGE-2 ↑	0.009	0.01	0.01	0.007
	ROUGE-L ↑	0.02	0.03	0.03	0.03
	BERTScore-P ↑	0.89	0.89	0.85	0.89
	BERTScore-R ↑	0.90	0.92	0.87	0.87
	BERTScore-F1 ↑	0.89	0.91	0.87	0.87
	SummaC ↑	0.61	0.64	0.66	0.65
	SARI ↑	40.2	40.1	37.6	37.7
	Flesch-Kincaid Grade Level (FKGL) ↓	14.7	10.2	13.7	14.3
	Dale-Chall Readability Score (DCRS) ↓	7.48	8.69	8.56	7.99
+ Reflection	ROUGE-1 ↑	0.03	0.03	0.03	0.03
	ROUGE-2 ↑	0.01	0.007	0.009	0.006
	ROUGE-L ↑	0.02	0.02	0.03	0.02
	BERTScore-P ↑	0.89	0.89	0.85	0.86
	BERTScore-R ↑	0.90	0.91	0.87	0.89
	BERTScore-F1 ↑	0.89	0.90	0.86	0.92
	SummaC ↑	0.56	0.59	0.60	0.61
	SARI ↑	40.0	39.9	37.7	37.6
	Flesch-Kincaid Grade Level (FKGL) ↓	14.7	13.9	15.8	14.7
	Dale-Chall Readability Score (DCRS) ↓	8.03	8.38	8.26	7.99

use though, factuality and aspect-specific summarization need metrics that better align with human intuition.

A. Limitations

There are a few limitations to this current work. While plain language summaries are generally intended for laypeople (i.e., people outside of academia), the PLS used in our work more align with academics reading a paper outside of their particular specialty rather than for laypeople. Also, we were unable to test other models and architectures due to time constraints. In terms of evaluation metrics, we use a variety of automated metrics that are commonly used for summarization, however, none of these metrics consider how well the summary adheres to the task specifically. Future work should focus on establishing metrics for individual tasks based on the particular task. Unfortunately, there are no task-specific automated summarization metrics currently. These will have to be developed if we are to properly evaluate aspect-based summarization. Finally, we currently only use automated metrics, which may not perfectly align with judgments made by human evaluators as previously mentioned [26]. Incorporating human evaluation into our metrics would significantly enhance this work.

VII. CONCLUSION

Automated summarization utilizing large language models (LLMs) enables the extraction and synthesis of crucial

information from longer source documents, converting them into brief summaries. This is especially beneficial in fields such as biomedicine, where clinicians and researchers require rapid access to key findings without needing to read entire articles, which can be rather time consuming. Personalized summarization further enhances this process by customizing summaries to meet individual preferences or specific professional requirements. In this study, we utilize four distinct author-generated summaries from biomedical review articles to create a biomedical dataset for aspect-based summarization. Moreover, we develop a baseline model with task-specific prompts tailored to each summary type and evaluate these using a comprehensive set of metrics. Overall, we find that using personalization and reflection appear to improve the factuality of the summaries, while maintaining similar, if not better, relevance and readability metrics.

While promising, future work needs to perform more extensive experimentation as well as expand the evaluation metrics. Future experiments should be done analyzing the frameworks on LLMs of different sizes (e.g., 70B+ parameter models), different LLMs (GPT-4, etc.), and modify/isolate changes to the architecture (e.g., persona and reflection steps). Additionally, the optimal evaluation metrics for this task remain uncertain. Expanding our factuality metrics to include AlignScore [10] may be a good start, but to most effectively align our evaluation with humans, we need to manually evaluate summaries.

VIII. Data Availability

All data and code used in the generation and evaluation of models are available at https://github.com/jomenke/Biomed-Review-Summarization. This repository contains links to the associated data, which are freely available through PMC-OAI, as well as discussion of the parameters used in our initial experiments.

References

[1] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[2] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008. [Online]. Available: http://arxiv.org/abs/1706.03762

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "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*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[5] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[6] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert systems with applications*, vol. 165, p. 113679, 2021.

[7] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating LLM hallucination via self reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1827–1843. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.123

[8] M. T. Hicks, J. Humphries, and J. Slater, "Chatgpt is bullshit," *Ethics and Information Technology*, vol. 26, no. 2, p. 38, 2024.

[9] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 163–177, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.10

[10] Y. Zha, Y. Yang, R. Li, and Z. Hu, "AlignScore: Evaluating factual consistency with a unified alignment function," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 11 328–11 348. [Online]. Available: https://aclanthology.org/2023.acl-long.634

[11] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "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)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 615–621. [Online]. Available: https://aclanthology.org/N18-2097

[12] G. Thomas, R. D. Hartley, and J. P. Kincaid, "Test-retest and inter-analyst reliability of the automated readability index, flesch reading ease score, and the fog count," *Journal of Reading Behavior*, vol. 7, no. 2, pp. 149–154, 1975.

[13] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educational research bulletin*, pp. 37–54, 1948.

[14] Y. Guo, T. August, G. Leroy, T. Cohen, and L. L. Wang, "Appls: Evaluating evaluation metrics for plain language summarization," *arXiv preprint arXiv:2305.14341*, 2023.

[15] T. Goldsack, C. Scarton, M. Shardlow, and C. Lin, "Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles," in *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, and J. Tsujii, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 122–131. [Online]. Available: https://aclanthology.org/2024.bionlp-1.10

[16] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv:2004.05150*, 2020.

[17] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "BioBART: Pretraining and evaluation of a biomedical generative language model," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 97–109. [Online]. Available: https://aclanthology.org/2022.bionlp-1.9

[18] Y. Guo, J. C. Chang, M. Antoniak, E. Bransom, T. Cohen, L. L. Wang, and T. August, "Personalized jargon identification for enhanced interdisciplinary communication," *arXiv preprint arXiv:2311.09481*, 2023.

[19] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[20] I. Titov and R. McDonald, "A joint model of text and aspect ratings for sentiment summarization," in *proceedings of ACL-08: HLT*, 2008, pp. 308–316.

[21] H. Hayashi, P. Budania, P. Wang, C. Ackerson, R. Neervannan, and G. Neubig, "Wikiasp: A dataset for multi-domain aspect-based summarization," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 211–225, 2021.

[22] X. Yang, K. Song, S. Cho, X. Wang, X. Pan, L. Petzold, and D. Yu, "Oasum: Large-scale open domain aspect-based summarization," *arXiv preprint arXiv:2212.09233*, 2022.

[23] S. Takeshita, T. Green, I. Reinig, K. Eckert, and S. Ponzetto, "ACLSum: A new dataset for aspect-based summarization of scientific publications," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh,

H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 6660–6675. [Online]. Available: https://aclanthology.org/2024.naacl-long.371

[24] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[25] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 401–415, 2016.

[26] B. Fang, X. Dai, and S. Karimi, "Understanding faithfulness and reasoning of large language models on plain biomedical summaries," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 9890–9911.

[27] S. Ramprasad and B. C. Wallace, "Do automatic factuality metrics measure factuality? a critical evaluation," *arXiv preprint arXiv:2411.16638*, 2024.

[28] Y. Guo, W. Qiu, Y. Wang, and T. Cohen, "Automated lay language summarization of biomedical scientific reviews," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 160–168.