IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 November 2024, accepted 13 December 2024, date of publication 17 December 2024,
date of current version 26 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3519095

RESEARCH ARTICLE

REMDoC: Reference-Free Evaluation for Medical Document Summaries via Contrastive Learning

JIMIN LEE, INGEOL BAEK, AND HWANHEE LEE, (Member, IEEE)
Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Hwanhee Lee (hwanheelee@cau.ac.kr)

This work was supported in part by the Chung-Ang University Research Grants, in 2023; and in part by the Institute for Information & Communications Technology Planning & Evaluation (IITP) through Korean Government (MSIT) through the Artificial Intelligence Graduate School Program, Chung-Ang University, under Grant 2021-0-01341.

ABSTRACT Despite significant advancements in automatic summary evaluation metrics based on pre-trained language models, accurately assessing the quality of medical document summaries remains a considerable challenge. Existing evaluation metrics often struggle to provide reliable assessments for medical summaries, particularly in the absence of reference texts. In this paper, we propose novel reference-free medical document summary evaluation metric, REMDoC: **R**eference-free **E**valuation for **M**edical **Do**cument Summaries via **C**ontrastive Learning which does not require reference summaries to evaluate summaries. REMDoC employs contrastive learning using medical text-tailored data augmentation techniques. Our primary motivation is to improve the alignment of automatic evaluations with human judgments, making the evaluation process more reliable and closer to medical expert assessments. Our research showcases the metric's superior performance in assessing the quality of generated summaries without the need for comparison texts. Through extensive experimentation and analysis, this work makes significant strides in improving the reliability and usability of automatic medical document evaluation tools in medical document settings.

INDEX TERMS Medical document summarization, contrastive learning, summary evaluation, reference-free evaluation.

I. INTRODUCTION

In the rapidly evolving field of healthcare systems, the ability to quickly and accurately summarize medical documents can significantly aid professionals in keeping abreast of the latest developments and making informed decisions. Medical documents are filled with specialized terminology, abbreviations, and jargon that can be challenging to interpret and summarize accurately [1], [17]. Moreover, medical documents often contain nuanced information that requires a professional understanding of the context. Hence, evaluating the quality of medical summaries is more complex than evaluating general document summary tasks. Also, medical document summary evaluation models struggle to capture medical details or nuances [14]. Because of these difficulties, researchers often require medical experts to assess the accuracy, completeness, and coherence of the summaries, making the evaluation process time-consuming and costly. For previous work on medical document summarization evaluation, widely used metrics such as ROUGE [11], BERTScore [27], and Delta-Ei [5] have been used. Some metrics leverage sentence-bert [19] and compute the cosine similarity between embeddings of target summary and generated summary to use as an evaluation metric. Numerous efforts have been made to identify and summarize the key points of related medical documents and to assess the quality of these summaries [8]. However, our findings demonstrate these metrics are not well-suited for medical document summarization evaluation, as n-gram similarity metrics such as ROUGE and models fine-tuned on datasets like SciFact [22] have disappointingly low correlations with human evaluation scores proposed in the previous work [24], under 0.053 as shown in Table 1.

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

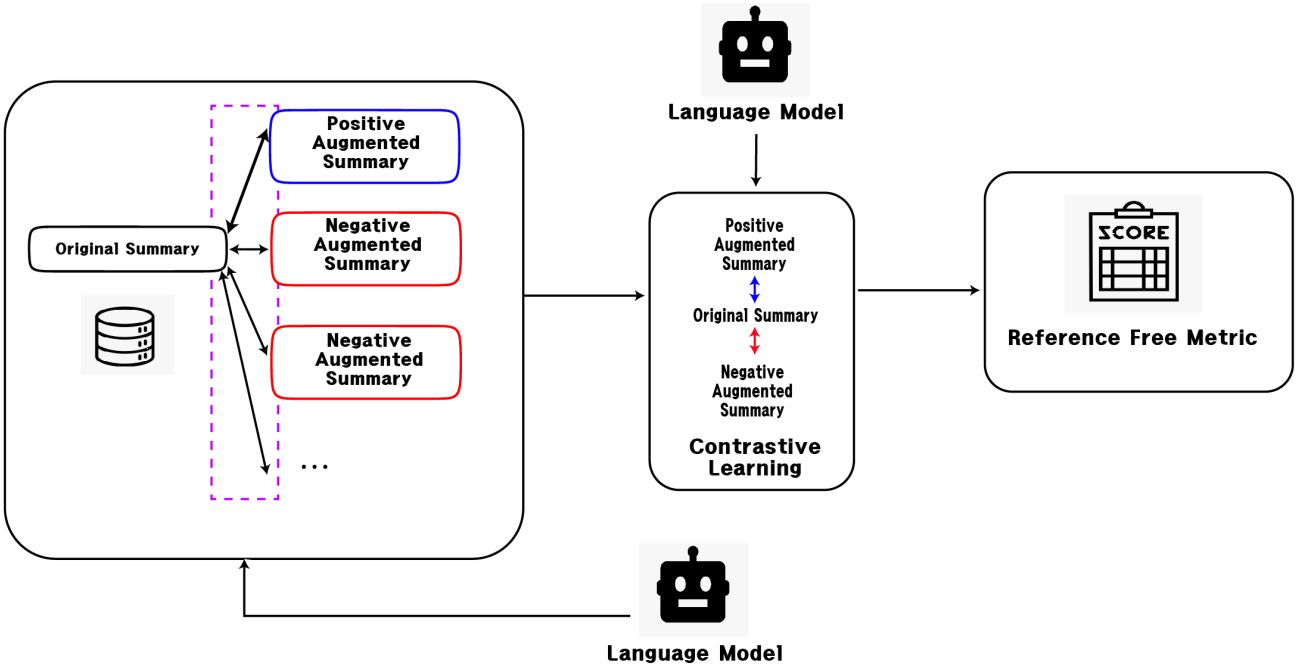

FIGURE 1. Overall training procedure of the reference-free medical document summary evaluation metric.

TABLE 1. Correlation coefficients with human judgments for the widely used metrics on medical document summarization tasks.

Method	ROUGE	BERT-S	NLI
ρ	-0.010	0.022	0.053

This suggests traditional metrics do not adequately reflect human judgment for summarization tasks. Due to the nature of medical text data, the meaning of a sentence can completely change with just one-word alteration, and the relationships between words are crucial. Capturing these minor changes is significantly important in evaluating the medical summaries. In this work, we propose a reference-free medical document summary evaluation metric that is not dependent on the reference ground truth medical summaries and use a document-summarization pair to evaluate the quality of medical document summaries. We develop our method upon the RoBERTa-large [13] model. We fine-tune the model via contrastive learning, where the model is trained to distinguish between the ground-truth summaries and precisely augment various positive and negative medical summaries. The model learns to differentiate the real summary from the augmented summary, enhancing its ability to accurately understand and evaluate summary like medical professionals. We evaluate our proposed metric on the human evaluation medical dataset. Our experimental results demonstrate high correlations with human evaluations, which outperform previous medical summary evaluation metrics.

II. RELATED WORK

EDA [25] suggests simple yet effective data augmentation techniques for the texts. EDA demonstrates various data augmentation techniques, such as synonym replacement, random swap, and random deletion methods, significantly enhancing text classification task performance. Named entity replacements significantly affect the performance of language models, highlighting their sensitivity to entity modifications [6]. Our research integrates and improves these augmentation techniques to help models capture the nuanced characteristics inherent in medical documents. Additionally, previous work [24] reveals a notable discrepancy between automated metrics and human evaluations in medical document summarization. They introduce a dataset of human-assessed summary quality facets from the medical document summarization for literature review and find out the inefficiency of conventional metrics for summarization evaluation. Our work leverages these human evaluation scores to substantiate our model's efficacy. Recent developments in summary evaluation emphasize improved alignment with human judgment by focusing on semantic and contextual understanding. SummScore [12] overcomes the limitations of traditional token-based metrics by using a cross-encoder model to evaluate semantic similarity directly with the original text. The Dense Passage Retrieval [7] approach is trained with both positive passages and negative passages, reinforcing the model's ability to distinguish relevant passages. Furthermore, UMIC [10] introduces a metric that does not require reference captions to evaluate image captions. By utilizing negative captions and fine-tuning the

UNITER model [2] through contrastive learning, this metric provides a robust evaluation framework. Drawing inspiration from these methodologies, we adopt a contrastive learning approach to train a metric for medical document summaries. By combining diverse data augmentation techniques with this contrastive learning framework, we propose a novel reference-free metric specifically designed for the evaluation of medical document summaries.

III. METHODS

We introduce a new medical document summarization metric REMDoC, **R**eference-free **E**valuation for **M**edical **Do**cument Summaries via **C**ontrastive Learning, through the following two steps. First, we augment the MSLR Cochrane [23] dataset using six different methods. By employing these augmentation techniques, we aim to accurately differentiate between summaries that retain the original meaning and those that do not. We formulate the final training dataset to this form: *(original summary, augmented summary)* (Sec III-A). Finally, we train the metric as a contrastive learning approach by learning the representation of medical document summaries by comparing similar *(positive medical summary)* and dissimilar *(negative medical summary)* pairs of data points (Sec III-B).

A. MEDICAL SUMMARY AUGMENTATION

Our augmentation approach is designed not only to introduce conspicuous differences but also to incorporate subtle, biomedically relevant variations from the original medical summaries. These methods are meticulously crafted to generate synthetic examples of negatively augmented summaries, thereby enriching the dataset for contrastive learning. Our method lies in using pairs of *(original summary, augmented summary)* as inputs for contrastive learning. The original summary is a concise representation of medical documents, while the augmented summary is a modified version of the original, incorporating augmented sentences aimed at enhancing the richness of the training data for contrastive learning. We choose the positive augmentation methods (synonym replacement, paraphrasing) because they retain expressions that are almost identical to the original summary, ensuring that the essential meaning remains unchanged. This helps the model learn to recognize valid yet slightly altered representations of the same medical content. On the other hand, the negative augmentation methods are designed to introduce more diversified and nuanced modifications that deviate from the original summary's content. These techniques (random deletion, random swap, antonym replacement, NER swap) create summaries with significant differences, such as the removal or distortion of critical information. This enables the model to distinguish between accurate summaries and those that contain misleading or incomplete information, further enhancing its ability to evaluate medical document summaries effectively. In this work, we utilize the following data augmentation techniques to generate new data pairs as in Table 2 for training the proposed metric. Our final training dataset consists of 22,350 augmented summaries derived from the 3,725 original summaries.

1) SYNONYM REPLACEMENT

Synonym Replacement (SR) is designed to generate sentences that are semantically similar by replacing certain words with their synonyms. This method helps us to diversify the linguistic expression within our dataset without deviating from the original meaning. The words are from WordNet lexical databases [15], which have words to change. For each selected word, look up synonyms that fit the context of the sentence. The next step is replacing the original words with their synonyms. For instance, we substitute 'low-quality' with 'inferior-quality,' 'comparing' with 'contrasting,' and 'advanced' with 'progressed.'.

2) PARAPHRASING

Paraphrasing (PAR) rephrases sentences to add structural diversity. We employ the Pegasus model [26], which shows the state-of-the-art performance in the paraphrasing multiple dataset task. This approach involves rephrasing sentences to create semantically similar but structurally distinct variants. By doing so, we enhance the generalizability and comprehension capabilities of our summarization model. This paraphrased sentence maintains the core meaning of the original but is reconstructed with different vocabulary and grammatical structures, which contributes to a richer environment for our summarization evaluation metric.

3) RANDOM DELETION

Random Deletion (RD) randomly removes words to simulate different forms of summarization compression. For each sentence in the dataset, we randomly select words for deletion, where each word has an equal chance of being deleted. We remove these selected words from the sentence.

4) RANDOM SWAP

Random Swap (RS) randomly selects pairs of words within a sentence and swaps their positions, which can increase linguistic diversity without significantly altering the semantic integrity of the information conveyed. In the original sentence, two specific swaps are performed. The term 'ovarian' was swapped with 'standard', and the terms 'women' and 'we' are also exchanged.

5) ANTONYM REPLACEMENT

Antonym Replacement (AR) is a data augmentation technique that replaces specific words or expressions in text with words or expressions with the opposite meaning. This approach aids our model in improving its ability to understand and distinguish opposite concepts. The first step involves identifying words within a sentence that can be replaced with their antonyms. The target word is replaced with its

TABLE 2. Examples of generated medical document summaries through the proposed data augmentation approaches.

Original Summary: We found only low-quality evidence comparing ultra-radical and standard surgery in women with advanced ovarian cancer and carcinomatosis.
Data Augmentation Approaches
Positive Augmentation Methods
Synonym Replacement (SR): We discovered solely inferior-quality proof contrasting extreme and conventional surgery in females with progressed ovarian cancer and carcinomatosis.
Paraphrase (PAR): Our research yielded only substandard evidence when evaluating the effectiveness of extreme versus traditional surgical approaches in females diagnosed with severe ovarian cancer and carcinomatosis.
Negative Augmentation Methods
Random Deletion (RD): We found low-quality evidence comparing radical and standard surgery in women with ovarian cancer.
Random Swap (RS): Women found only low-quality evidence comparing ovarian and ultra-radical surgery in we with carcinomatosis and advanced standard cancer
Antonym Replacement (AR): We found only high-quality evidence comparing conservative and advanced surgery in women without early ovarian cancer and carcinomatosis.
Named Entity Replacement (NER): They found only middle-quality proof comparing ultra-radical and standard surgery in children with advanced ovarian torsion and carcinomatosis

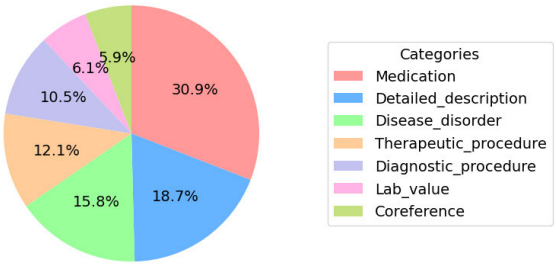

FIGURE 2. Distribution of entity classes in the MSLR dataset.

antonym. For example, we replaced 'with' with the antonym of 'without' as in Table 2.

6) NAMED ENTITY REPLACEMENT

Named Entity Replacement (NER) method substitutes named entities with others of the same category to enrich the model's ability to handle factual information. In our study, we implemented a novel approach to data augmentation for the Cochrane dataset's training set by leveraging the *d4data/biomedical-ner-all*[1] model, which recognizes 84 biomedical entities. This pre-trained model identifies and categorizes biomedical entities into various entity group categories such as *lab value*, *detailed description*, *therapeutic procedure*, *disease disorder*, *medication*, *diagnostic procedure*, and *sign symptom*. Upon analyzing the training dataset, we gathered words belonging to these seven distinct entity groups. The augmentation process involved swapping entities within the same category across different data instances. For instance, if 'headache' is labeled as a Sign symptom in the first training entry and 'nausea' is also labeled as a Sign symptom in the fifth, we swapped these two terms to create a new, augmented training data instance. This method of intra-group entity swapping aims to enrich the dataset by diversifying the context in which each term is used, potentially improving the robustness of models trained on this augmented dataset. The swapping technique is carefully designed to maintain the integrity of the medical context, ensuring that the swapped entities are contextually appropriate. This augmentation strategy not only augments the size of the training data but also introduces a level of variance that can enhance the generalization capabilities of the model.

[1]https://huggingface.co/d4data/biomedical-ner-all

B. CONTRASTIVE LEARNING

Contrastive learning is a method that extracts features by minimizing the representation distance between similar samples and maximizing the distance between representations of different samples. By employing contrastive learning, we can train the metric without needing specific labels, thereby reducing the effort required from human annotators. We construct a dataset based on the original summary (q_i) and positive augmentation methods, SR and PAR, and negative

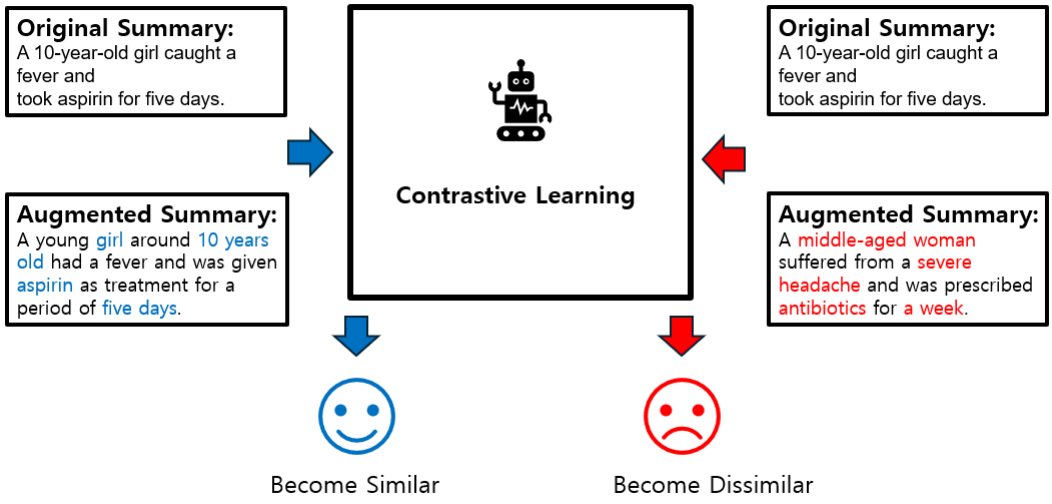

FIGURE 3. Proposed contrastive learning framework for fine-tuning a metric.

Algorithm 1 Flow of NER Swap Algorithm

Data: Set of documents $\mathbf{D} = \{d_k\}_{k=1}^{T}$
Set of NERs $\mathbf{N} = \{n_{i,k}\}_{i,k=1}^{L,T}$
Result: NER swapped Medical Summaries $\mathbf{S}^*$
$\mathbf{S}^* \leftarrow []$; ; // Initialize the output list
for $k = 1$ **to** T **do**
  for $j = 1$ **to** len$(NER_{d_{:,k}})$ **do**
    $u \leftarrow$ Random$(N_{j,T})$; ; // Select a random number u from $N_{j,T}$
    d_k.Replace$(NER_{d_{j,k}}, n_{j,u})$; ; // Replace NER in document
    $\mathbf{N}$.Remove$(n_{j,u})$; ; // Remove used NER
  $\mathbf{S}^*$.Append(d_k); ; // Append modified document to list
return $\mathbf{S}^*$

augmentation methods, RD, RS, AR, and NER.

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^{m} \tag{1}$$

$\mathcal{D}$ is the training data that consists of m instances. We utilize an in-batch negative function for positive samples (p_i^+) and negative samples (p_i^-). During training, each of the two positive samples is included separately, resulting in the creation of its own loss function. The mathematical formulation of the L_{cs} is as follows:

$$L_{cs}(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-) = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}}, \tag{2}$$

where $\text{sim}(q_i, p_i^+)$ is a cosine similarity between q_i and p_i^+. This approach intuitively aligns with the goal of medical document summary evaluation. For positive augmented datasets, our model learns to minimize the distance between their representations, recognizing them as different expressions of the same fundamental information. On the other hand, if a negative augmented summary diverges significantly, by introducing unrelated information or omitting critical details, our model increases the representational distance, highlighting the loss or distortion of information. Through this approach, the model can optimize both learning objectives simultaneously. The training process does not solely depend on the static training data but rather on the model's capacity to minimize this contrastive loss by updating its parameters. These parameter updates enable the model to generalize and perform well on unseen data. Thus, although the loss function utilizes training data pairs during optimization, it inherently drives the learning of representations that capture more generalizable relationships within the data, going beyond the specific samples in the training set.

IV. EXPERIMENTS

A. BASELINES

We employ a facet-based human evaluation methodology known as PIO [18]. PIO alignment stands for population, intervention, and outcome as the human evaluation score. Population metric assesses the degree to which the population described in the generated summary aligns with that in the target summary. Intervention evaluates the consistency of the intervention details between the generated and target summaries. Furthermore, we employ fluency, direction, and strength scores which are also important alignment metrics for evaluating the quality of summaries. The outcome measures the concordance of the outcomes reported in the generated summary with those in the target summary.

We measure the correlation between human evaluation scores and metric scores. For traditional summary evaluation metrics, we utilize ROUGE [11], BERTScore [27], Delta-EI [4], [23], NLI and ClaimVer [3]. ROUGE computes the n-gram similarity scores, while BERTScore assesses the generated summaries using the BERT model. Lastly, Delta-EI measures the probability distributions of evidence direction among intervention-output pairs of the target and generated summary. For the cases of NLI and ClaimVer, the procedure follows in the same manner as described in the previous work [24]. We also utilize Large Language Model (LLM) prompts as baselines to score the similarity between original summaries and model-generated summaries as in Table 3. We utilize BioMistral-7B [9], which is an open-source pre-trained LLM for medical domains. Llama3-8B,[2] Llama3.1-8B[3] and Orca-7B [16] are state-of-the-art open-source Large Language Models from Meta and Microsoft. Using the output scores, we compare them with PIO scores to analyze their correlations like other metrics.

TABLE 3. Prompt used for evaluating summaries using LLM.

Prompt
Score the following summary with respect to population, intervention, and outcome on a continuous scale of 0 to 1, where a score of zero means the summary does not entail the original meaning of the document, and a score of one means the summary entails the original meaning of the document. [Medical Document] [Summary] Score:

B. IMPLEMENTATION DETAILS

For the paraphrasing in the data augmentation step, we utilize the pre-trained Pegasus model for paraphrasing [26]. We employ the MSLR Cochrane dataset [21] and the augmented datasets to train the model. In our comprehensive study, we conduct a detailed comparison between human evaluation scores and embeddings generated by models for summaries. To achieve this, we utilize RoBERTa-base, DistillRoBERTa-base [20], and RoBERTa-large [13] as our models. These models embed text to measure the similarity between the original summary and the augmented summary. For the test dataset, we utilized the MSLR Cochrane dataset, which comprises 597 sets (ground truth summary and model-generated summary) from 6 models and their PIO, fluency, direction, and strength scores. We compare the PIO scores and REMDoC scores using the Spearman correlation (ρ). We configure the batch size of 128 and train for 4 epochs with a learning rate of 10^{-4}, using the AdamW optimizer.

[2]https://huggingface.co/meta-llama/Meta-Llama-3-8B
[3]https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

C. COMPUTATIONAL RESOURCES

We use AMD Ryzen 5 5600G CPU (3.90 GHz) with two NVIDIA RTX 4090 GPUs for the experiments. The software environments are Python 3.11.5 and PyTorch 2.3.1.

TABLE 4. Results of correlation coefficients between automated metrics and PIO, Fluency, Direction and Strength.

Metric	PIO ρ	Fluency ρ	Direction ρ	Strength ρ
Rouge	-0.010	-0.014	0.007	-0.035
BERTScore	0.022	0.000	0.036	-0.033
Delta-EI	-0.080	0.066	-0.006	0.054
STS	-0.042	-0.042	0.001	-0.056
ClaimVer	0.142	-0.051	-0.017	-0.093
NLI	0.053	-0.026	-0.011	-0.063
BioMistral-7B	0.069	-0.026	0.021	-0.012
Llama3-8B	0.140	0.031	0.125	-0.072
Llama3.1-8B	0.287	**0.386***	0.291	0.263
Orca-7B	0.131	-0.016	-0.010	0.123
REMDoC				
RoBERTa-base	0.415	0.291	0.387	0.231
RoBERTa-large	**0.519***	0.312	**0.401***	**0.491***
Distill-RoBERTa	0.357	0.243	0.342	0.212

D. PERFORMANCE COMPARISON

we present comprehensive experimental results in Table 4 that highlight significant improvements in the correlation between automated metrics and human evaluations through the use of various data augmentation strategies combined with contrastive learning. Our findings emphasize the crucial role of data augmentation techniques and model scale in achieving superior performance. Notably, the RoBERTa-large model, when subjected to a combination of all six augmentation methods, achieves the highest correlation coefficient of 0.519. In terms of fluency, the REMDoC model shows second highest among the models compared. Llama 3.1-8B model achieved the highest scores in the fluency evaluation. This outcome suggests that the inherent knowledge embedded within large language models (LLMs) enables them to more accurately assess and measure the fluency of summaries. Additionally, REMDoC demonstrates a robust correlation of 0.401 for direction, indicating its effectiveness in modeling directional aspects of language that align well with human judgment. For the strength aspect, REMDoC achieves a correlation of 0.491, again outperforming other models examined. This exceptional performance can be attributed to the model's larger capacity, enabling it to better process and integrate complex variations in training data, resulting in assessments that closely mirror human evaluations. Additionally, we observe that all three models, RoBERTa-base, RoBERTa-large, and Distill-RoBERTa-base, exhibited superior performance when trained on augmented data compared to their original metrics. This improvement indicates that data augmentation not only enhances the robustness and generalization capabilities of the models but also significantly boosts their alignment with human evaluations. Furthermore, We visualize the

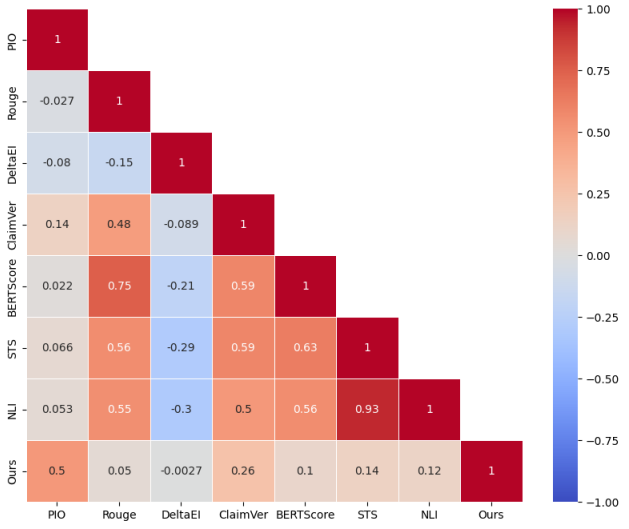

FIGURE 4. Correlation coefficients between various metrics including widely used metrics and REMDoC.

correlation between previous metrics and the proposed metric as shown in Figure 4. Human evaluation (PIO) displays weak correlations with metrics except for our proposed metric. STS and NLI show a high correlation, which leverages sentence bert. Our metric exhibits weak correlations with most metrics except for DeltaEI, suggesting DeltaEI captures somewhat similar aspects of evaluation. We find that correlations for the LLM-based evaluations are especially low compared to other methods. Our findings reveal that despite the application of recent LLM prompt engineering techniques, it remains challenging to produce results that closely resemble human evaluation.

E. ABLATION STUDY

1) PERFORMANCE AMONG EACH DATA AUGMENTATION APPROACH

For the ablation study, we conduct experiments by measuring the correlation each time we remove a specific augmentation method. In Table 5, we observe that removing positive samples (Synonym Replacement (SR) and Paraphrase (PAR)) leads to a decrease in correlation scores. Additionally, we notice the most significant score drop when removing Antonym (AR) and NER Swap (NER). In the Roberta large model training, removing AR and NER individually results in a decrease of 0.124 and 0.145, respectively. These insights collectively point to the efficacy of targeted data augmentation in bridging the gap between automated evaluations and human perceptions, suggesting a path forward for enhancing the reliability and human-likeness of model outputs. The nuanced understanding of text provided by these augmentation methods fosters a deeper comprehension of context, which is essential for models tasked with evaluating text in a manner that resonates with human interpretations.

TABLE 5. Results of correlation coefficients between automated metrics and human evaluation(PIO) across different data augmentation methods.

Models	ρ
REMDoC-RoBERTa-base	0.415
$w/o\ SR\ (p^+)$	0.343
$w/o\ PAR\ (p^+)$	0.367
$w/o\ RD\ (p^-)$	0.358
$w/o\ RS\ (p^-)$	0.354
$w/o\ AR\ (p^-)$	0.319
$w/o\ NER\ (p^-)$	0.324
REMDoC-RoBERTa-large	**0.519**
$w/o\ SR\ (p^+)$	0.453
$w/o\ PAR\ (p^+)$	0.433
$w/o\ RD\ (p^-)$	0.407
$w/o\ RS\ (p^-)$	0.431
$w/o\ AR\ (p^-)$	0.395
$w/o\ NER\ (p^-)$	0.374
REMDoC-Distill-RoBERTa-base	0.357
$w/o\ SR\ (p^+)$	0.343
$w/o\ PAR\ (p^+)$	0.329
$w/o\ RD\ (p^-)$	0.308
$w/o\ RS\ (p^-)$	0.341
$w/o\ AR\ (p^-)$	0.298
$w/o\ NER\ (p^-)$	0.301

TABLE 6. Performance comparison according to the number of NER groups.

Models	NER Category	ρ
REMDoC-RoBERTa-base	1	0.420
	3	0.430
	5	0.435
	7	0.415
REMDoC-RoBERTa-large	1	0.461
	3	0.468
	5	0.466
	7	**0.519***
REMDoC-Distill-RoBERTa-base	1	0.414
	3	0.419
	5	0.420
	7	0.357

2) PERFORMANCE AMONG THE NUMBER OF NER SWAP

For each model, we evaluate the correlation performance after swapping 1, 3, 5, and 7 NER categories as shown in Table 6. For RoBERTa-base and DistilRoBERTa-base, the performance generally improves as the number of swapped NER categories increases. This trend suggests that these models can adapt to the introduced variability up to a certain extent. The most significant improvement is observed in RoBERTa-large with 7 NER swaps, achieving a correlation coefficient of 0.519, highlighting its superior capacity to handle extensive NER category swaps.

F. CASE STUDY

As illustrated in Table 7, the document and summary exhibit nearly identical content, indicating a very high correlation between them. However, existing metrics such as ROUGE-1, BERTScore, and NLI report low correlation coefficients

TABLE 7. Case study on evaluating the original summary, exaggerated summary and summary with irrelevant details.

Document
The clinical features of 49 patients who had sustained small strokes in the internal carotid artery territory, who were normotensive, free from cardiac or other relevant disease, and who each had a normal appropriate single vessel angiogram are presented. These were randomized into two groups: group A, 25 patients, who received only supportive treatment; group B, 24 patients who were treated with anticoagulants for an average period of 18 months. There was a reduced incidence of neurological episodes during the administration of anticoagulant therapy but, after treatment was discontinued, there was no significant difference between the two groups. In view of the relatively benign prognosis for this syndrome, unless special facilities exist for the personal control of anticoagulant treatment, the dangers may outweigh the benefits.
Summary 1 (Original summary)
Compared with control, there was no evidence of benefit from long-term anticoagulant therapy in people with presumed non-cardioembolic ischaemic stroke or transient ischaemic attack, but there was a significant bleeding risk.
Summary 2 (Exaggerated summary)
A study of 49 patients with small strokes in the internal carotid artery territory highlighted the striking benefits of anticoagulant therapy. The group treated with anticoagulants showed a dramatic reduction in neurological episodes, with effects persisting even after treatment ended. However, the potential risks of this powerful therapy require careful management, as the dangers could outweigh the benefits without specialized oversight.
Summary 3 (Summary with irrelevant details)
In a study of 49 patients with small strokes in the internal carotid artery territory, those treated with anticoagulants for 18 months—interestingly the same amount of time needed to train for a marathon—showed a significant reduction in neurological episodes. Despite their pristine angiograms, which are as rare as a four-leaf clover, concerns were raised that without specialist oversight, the risks of anticoagulant therapy might outweigh the benefits, much like the delicate care needed for a bonsai tree.

Metric	**Summary 1 Score**	**Summary 2 Score**	**Summary 3 Score**
REMDoC	0.824	0.412	0.571
ROUGE-1	0.226	0.267	0.319
BERTScore	0.463	0.856	0.885
NLI	0.340	0.721	0.561
Llama3.1-8B	0.800	0.200	0.500

between the document and summary, failing to recognize their semantic overlap accurately. In contrast, our proposed metric successfully evaluates the high correlation between the document and the summary. By accurately capturing the degree to which the summary preserves the document's original meaning, our metric demonstrates its effectiveness in assessing summary quality. This underscores the superiority of our metric over traditional approaches in evaluating medical document summaries, particularly in cases where semantic preservation is critical. In addition to the experiments presented in Table 7, we conducted further evaluations using summaries that included exaggerated information and irrelevant details. Interestingly, while our proposed metric demonstrated strong performance in assessing the quality of summaries that closely preserved the original document's meaning, the Llama 3.1-8B model excelled in identifying and penalizing these exaggerated and irrelevant summaries more accurately than our model, REMDoC. This improved performance of the Llama3.1-8B model can likely be attributed to the vast amount of inherent knowledge it possesses. With access to a broader context and a deeper understanding of language nuances, the Llama3.1-8B model is better equipped to detect discrepancies and ensure that the summaries maintain a high degree of relevance and accuracy. This highlights the strengths of large language models in capturing subtleties that might be missed by more specialized metrics, particularly in scenarios where summaries deviate from the original content in misleading ways.

V. CONCLUSION

This paper proposes a reference-free medical multi-document summary evaluation of the metric via contrastive learning to the pre-trained language models. Experimental results demonstrate that our evaluating model outperforms existing metrics, indicating strong alignment with human evaluations. We expect that our findings can be adapted and expanded, serving as a stepping stone toward broader medical applications.

VI. LIMITATIONS

Access to large-scale, diverse, real-world medical datasets is challenging due to privacy and regulatory limitations, which restrict the availability of publicly accessible, representative medical data. Additionally, while our augmentation techniques help address data size limitations, they may not perfectly capture the complexity and variability of true clinical scenarios, potentially impacting generalizability to diverse real-world applications. Future work could test the model in varied healthcare environments to further validate its robustness.

VII. ETHICAL CONSIDERATION

We recognize the potential implications of relying on automated systems in clinical settings, where decisions could significantly impact patient care. To mitigate risks, we recommend that REMDoC be used as a supplementary tool rather than a replacement for human judgment. Continuous monitoring and periodic evaluation of the metric's performance in real-world scenarios are necessary to prevent biases or inaccuracies from affecting clinical outcomes. Additionally, it is important to acknowledge that the datasets used for training and evaluating REMDoC may be inherently limited. These datasets are curated and evaluated by human experts, whose assessments can be subjective and influenced by personal biases. The limited diversity of these datasets may not fully capture the complexity of real-world clinical scenarios, potentially leading to skewed outcomes. Therefore, it is essential to continually expand and diversify the datasets, and to be aware of the limitations of human evaluation in order to refine REMDoC.

REFERENCES

[1] M. Abdollahi, X. Gao, Y. Mei, S. Ghosh, J. Li, and M. Narag, "Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques," *Artif. Intell. Med.*, vol. 120, Oct. 2021, Art. no. 102167.
[2] Y.-C. Chen. (2020). *Uniter: Universal Image-Text Representation Learning*.
[3] P. P. S. Dammu. (2024). *Claimver: Explainable Claim-Level Verification and Evidence Attribution of Text Through Knowledge Graphs*.
[4] J. DeYoung, I. Beltagy, M. v. Zuylen, B. Kuehl, and L. L. Wang, "MS2: Multi-document summarization of medical studies," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Jan. 2021, pp. 7494–7513.
[5] J. DeYoung, E. Lehman, B. Nye, I. Marshall, and B. C. Wallace, "Evidence inference 2.0: More data, better models," in *Proc. 19th SIGBioMed Workshop Biomed. Lang. Process.*, 2020, pp. 123–132.
[6] S. Goodarzi, N. Kagita, D. Minn, S. Wang, R. Dessi, S. Toshniwal, A. Williams, J. Lanchantin, and K. Sinha, "Robustness of named-entity replacements for in-context learning," in *Proc. Findings Assoc. Comput. Linguistics*, 2023, pp. 10914–10931.
[7] V. Karpukhin. (2020). *Dense Passage Retrieval for Open-Domain Question Answering*.
[8] B. Khan, Z. A. Shah, M. Usman, I. Khan, and B. Niazi, "Exploring the landscape of automatic text summarization: A comprehensive survey," *IEEE Access*, vol. 11, pp. 109819–109840, 2023.
[9] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "BioMistral: A collection of open-source pretrained large language models for medical domains," 2024, *arXiv:2402.10373*.
[10] H. Lee, S. Yoon, F. Dernoncourt, T. Bui, and K. Jung. (2021). *Umic: An Unreferenced Metric for Image Captioning via Contrastive Learning*.
[11] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, vol. 2, pp. 74–81, Jul. 2004.
[12] W. Lin, S. Li, C. Zhang, B. Ji, J. Yu, J. Ma, and Z. Yi. (2022). *Summscore: A Comprehensive Evaluation Metric for Summary Quality Based on Cross-encoder*.
[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[14] S. Meystre and P. J. Haug, "Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation," *J. Biomed. Informat.*, vol. 39, no. 6, pp. 589–599, Dec. 2006.
[15] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[16] A. Mitra, L. Del Corro, S. Mahajan, A. Codas, C. Simoes, S. Agarwal, X. Chen, A. Razdaibiedina, E. Jones, K. Aggarwal, H. Palangi, G. Zheng, C. Rosset, H. Khanpour, and A. Awadallah, "Orca 2: Teaching small language models how to reason," 2023, *arXiv:2311.11045*.
[17] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156043–156070, 2021.
[18] Y. Otmakhova, K. Verspoor, T. Baldwin, and J. H. Lau, "The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, May 2022, pp. 5098–5111.
[19] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.
[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[21] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021, pp. 1–26.
[22] D. Wadden, K. Lo, B. Kuehl, A. Cohan, I. Beltagy, L. L. Wang, and H. Hajishirzi, "SciFact-open: Towards open-domain scientific claim verification," 2022, *arXiv:2210.13777*.
[23] B. Wallace, S. Saha, F. Soboczenski, and I. Marshall, "Generating (factual) narrative summaries of RCTs: Experiments with neural multi-document summarization," in *Proc. AMIA. Annu. Symp. AMIA Symp.*, Jan. 2020, pp. 605–614.
[24] L. L. Wang. (2023). *Automated Metrics for Medical Multi-Document Summarization Disagree With Human Evaluations*.
[25] J. Wei and K. Zou. (2019). *Eda: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks*.
[26] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. (2020). *Pegasus: Pre-Training With Extracted Gap-sentences for Abstractive Summarization*.
[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. (2020). *Bertscore: Evaluating Text Generation With BERT*.

JIMIN LEE received the B.S. degree in statistical modeling data sciences from The Pennsylvania State University, University Park, USA, in 2022. He is currently pursuing the M.S. degree with the Department of Artificial Intelligence, Chung-Ang University. His research interests include summarization and NLP applications for the medical domain.

INGEOL BAEK received the B.S. degree in urban engineering from Jeonbuk National University, South Korea, in 2023. He is currently pursuing the M.S. degree with the Department of Artificial Intelligence, Chung-Ang University. His research interests include information retrieval and the factuality of language models.

HWANHEE LEE (Member, IEEE) received the B.S. and Ph.D. degrees in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2017 and 2022, respectively. He was a Research Intern with the NAVER AI Laboratory, from August 2021 to January 2022. He has been an Assistant Professor with the Department of Artificial Intelligence, Chung-Ang University, Seoul, since March 2023. His main research interest includes mitigating and detecting hallucination of language models.