Improving Medical Question Summarization through Re-ranking

Sibo Wei*, Xueping Peng†, Yan Jiang*(✉), Zhao Li*,¶, Yan Liu‡, Zhiqiang Wang§, Wenpeng Lu*(✉),
* Key Laboratory of Computing Power Network and Information Security, Ministry of Education,
Shandong Computer Science Center (National Supercomputer Center in Jinan),
Qilu University of Technology (Shandong Academy of Sciences);
Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing,
Shandong Fundamental Research Center for Computer Science, Jinan, China
† Australian Artificial Intelligence Institute, University of Technology Sydney, Australia
‡ Yingdong Intelligent Technology (Shandong) Co., Ltd., Jinan, China
§ China Mobile Group Shandong Co., Ltd., Jinan, China
¶ Evay Info Co., Ltd., Jinan, China
(✉) Corresponding author email: {jiangy, wenpeng.lu}@qlu.edu.cn

***Abstract*—Fine-tuning sequence-to-sequence (Seq2Seq) models applied on downstream datasets have gained remarkable success in the task of medical question summarization (MQS). However, Seq2Seq models suffer from a discrepancy between their objective function and evaluation metrics, whose objective function is based on local and token-level predictions whereas the evaluation metrics of MQS focus on overall similarity between the gold references and system predictions. To address the gap, it is crucial to consider multiple candidate summaries, assess their quality and re-rank them to obtain the optimal summary. In this paper, we propose to enhance MQS through re-ranking. Specifically, we introduce two re-rankers (SimGate and LLM) designed to perform reference-free evaluations on candidate summaries and select the best one. The former evaluates candidate summaries using text semantic similarity and a confidence gate, while the latter leverages the powerful comprehension capability of large language models to identify more preferable summaries. Experimental results demonstrate the effectiveness of the proposed re-ranking mechanisms. The SimGate re-ranker, compared with BART, exhibits a significant improvement in the ROUGE score on all datasets. Additionally, LLM re-rankers also demonstrate significant improvements on most datasets, underscoring the promising potential of large language models as re-rankers. The code and datasets are available at https://github.com/yrbobo/MQS-Reranker.**

***Index Terms*—Medical Question Summarization, Natural Language Generation, Question Answering, Re-ranker**

I. Introduction

The rise in popularity of online healthcare services has resulted in an increasing number of individuals turning to medical community websites in search of answers to their health-related queries [1]. However, the sheer volume of consumers and questions make it impractical to solely rely on manual responses from professionals. This necessitates the use of an automated medical question-answering system (MQAS). The questions submitted by consumers often contain redundant or irrelevant information, as well as unprofessional descriptions, which poses a challenge for MQAS in retrieving relevant answers. In order to provide accurate responses, MQAS must accurately comprehend the intent behind the questions [2], which is a crucial and challenging task. Various solutions have been proposed by researchers, including query relaxation [3], question entailment [4]–[6], and question summarization [2], [5]–[7]. Among these approaches, question summarization has shown the most promising results and attracted considerable attention.

TABLE I
Performance Comparison on CHQ-Summ Dataset. The candidate summaries are generated by a pre-trained BART model, and we select the best and the worst candidates (w.r.t. ROUGE scores) for each of the samples. **Best** and **Worst** represent the average performance of the best and worst candidates. **Improvement** is obtained by comparing the best re-ranker and the original BART model. R1/2/L refer to ROUGE-1/2/L scores.

System	R1	R2	RL
BART	42.40	24.00	39.79
Best	59.45	40.44	57.42
Worst	19.30	6.75	16.18
BART + LLM re-ranker (ours)	43.59	25.20	40.94
BART + SimGate re-ranker (ours)	**44.00**	**25.47**	**41.15**
Improvement(%)	3.77	6.13	3.42

The goal of medical question summarization (MQS) is to condense lengthy consumer health questions (CHQs) into concise frequently asked questions (FAQs). FAQs, unlike verbose CHQs, capture the core intent of the original questions by emphasizing essential information and eliminating unnecessary and redundant content. This facilitates MQAS in retrieving accurate answers more effectively.

The majority of existing methods for MQS rely on sequence-to-sequence (Seq2Seq) models, which are adapted and enhanced to meet the requirements of MQS. Popular Seq2Seq models used for MQS include Transformer [8], ProphetNet [9], PEGASUS [10], and BART [11], with BART achieving the highest performance [12]. Existing studies have

made efforts to enhance MQS from different perspectives. Some studies improve the generative transformer models by incorporating structured knowledge or lexical resources [13], [14]. Other works strengthen MQS using different machine learning strategies, including transfer learning [15]–[17], multi-task learning [5], [6], reinforcement learning [18] and contrastive learning [2], [7]. Although the existing work has significantly improved the performance of MQS, all of them are based on Seq2Seq models, leading to unavoidable challenges related to the flaws in Seq2Seq architecture.

All the aforementioned works mentioned above fail to address the mismatching between the learning objective of Seq2Seq models and evaluation metrics of MQS task [12], [19]. Specifically, Seq2Seq models are trained using the framework of maximum likelihood estimation (MLE), which relies on local and token-level predictions as the objective function. However, the evaluation metrics for MQS focus on the overall similarity between the gold references and the system predictions. The mismatch between the objective function and evaluation metrics implies that the unique summary generated by Seq2Seq may not necessarily be the best, while the candidate summaries through beam search may perform better in terms of MQS metrics. To bridge the gap, it is crucial to consider more candidate summaries, evaluate their quality and re-rank them to select the optimal summary. The performance of fine-tuned BART and our two proposed re-rankers is compared in Table I. BART, which is based on Seq2Seq framework, generates one single summary using beam search. On the other hand, our re-rankers maintain 16 candidate summaries and re-rank them to obtain the best one. According to the table, our re-rankers significantly outperform BART, demonstrating the necessary and effectiveness of re-ranking mechanism. While re-ranking has been applied in general abstractive summarization [12], [19], relatively limited work has been done on re-ranking for MQS due to its complexity and specificity.

To address this research gap, we propose a two-stage framework for MQS that incorporates two novel re-ranking mechanisms. Firstly, we fine-tune a Seq2Seq model to generate candidate summaries. Subsequently, we employ a re-ranker to perform a reference-free evaluation of the candidate summaries and select the most suitable one. Specifically, we introduce two re-rankers, namely SimGate and LLM. SimGate evaluates candidate summaries by considering text semantic similarity and includes a confidence gate. On the other hand, LLM leverages the powerful comprehension capability of large language models to identify superior summaries. Extensive experiments demonstrate the significant improvement achieved by our proposed re-rankers. These re-rankers are simple yet effective, enabling seamless integration into any existing framework as flexible components.

Accordingly, this paper makes the following major contributions:

- We introduce a two-stage framework for MQS task. In the first stage, a Seq2Seq model generates candidate summaries, while the second stage employs a re-ranker to evaluate the candidates and select the optimal summary. We propose two flexible re-rankers, SimGate and LLM, which improve the framework. This framework represents the first attempt to leverage re-ranking mechanisms to enhance the quality of medical question summarization in the MQS task.
- We propose SimGate, a novel re-ranker that employs text semantic similarity and a confidence gate to evaluate and re-rank the candidate summaries. Extensive experiments show that SimGate outperforms other methods, achieving a new state-of-the-art result in MQS.
- We propose LLM, a novel re-ranker that transforms the re-ranking into a multiple-choice understanding task, leveraging the comprehension capabilities of large language models (LLMs) to select better candidate summaries. Extensive experiments demonstrate the potential of LLMs as re-rankers, even in few-shot scenarios, suggesting promising prospects for their application in MQS.

II. Related Work

A. Medical Question Summarization

In recent years, the MQS task has gained increasing attention from industrial and research communities. The task was originally introduced by Ben Abacha et al. in 2019 [20]. They manually annotated the initial MQS dataset, MeQSum, and utilized pointer generation networks to generate summaries for consumer health questions. In 2021, Ben Abacha et al. organized the MEDIQA shared task [21], sparking renewed interest in MQS. In this shared task, some researchers focused to enhancing generative transformer models through the integration of structured knowledge or lexical resources. For example, Sang et al. integrated and enriched the structured knowledge into generative transformer models using medical entity embeddings [13]. He et al. utilized lexical resources to correct the errors generated by generative transformer models and devised a simple heuristic for re-ranking the outputs [14]. Other works explored the utilization of transfer learning or multi-task learning to improve MQS performance. For instance, Balumuri et al. [15], Mrini et al. [16], and Yadav et al. [17] applied transfer learning on MQS by leveraging the knowledge of pre-trained language models to improve the performance. Mrini et al. [5], [6] proposed a multi-task learning framework that jointly optimized medical question summarization and entailment tasks. Moreover, recent studies have recognized the importance of question focuses in MQS, as they directly influence the quality and reliability of the summary. To further enhance performance, researchers have employed reinforcement learning and contrastive learning. Yadav et al. introduced a reinforcement learning-based framework that utilized rewards derived from sub-tasks such as question-type identification and question-focus recognition [18]. Zhang et al. [2] and Wei et al. [7] employed contrastive learning to emphasize the focus of the question by utilizing the overlap phrases between CHQ and FAQ or medical entities [22], [23], generating hard negative samples and subsequently improving MQS performance.

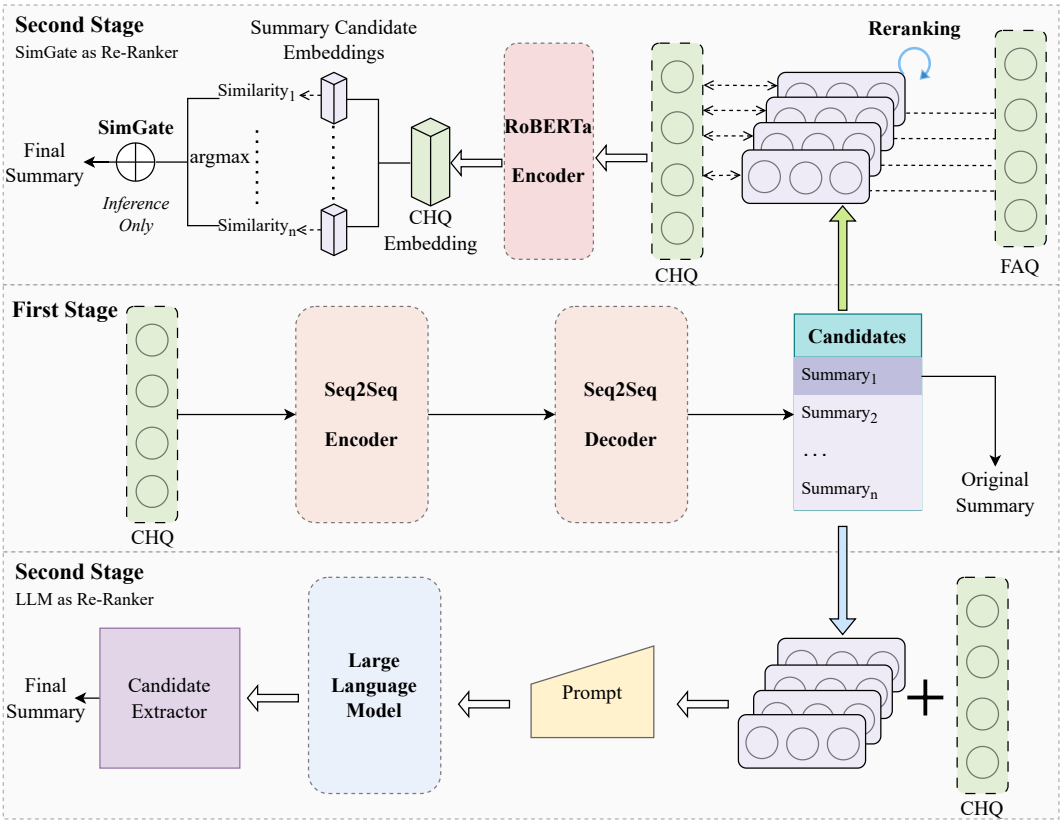

Fig. 1. The framework of our proposed model, which contains two stages, one for generating multiple candidate summaries with Seq2Seq models, another for re-ranking them to select the best one as final summary. We realize two re-rankers for the second stage. One is SimGate re-ranker, which utilizes text semantic similarity and a confidence gate to select the final summary from the candidate ones. Another is LLM re-ranker, which utilizes prompt template to guide large language model to select the best summary for the original questions.

Despite the achievement of the aforementioned works in MQS, it is important to note that all of them rely on the Seq2Seq architecture. However, this approach overlooks the issue of mismatching between the learning objective of Seq2Seq and the evaluation metrics [12], [19]. To address the gap, it becomes necessary to evaluate the quality of multiple candidate summaries, re-ranking them to select the optimal one. Surprisingly, there is currently no related work in the MQS task specifically aimed at resolving the mismatching problem. This highlights the need for researchers to devote more attention to MQS.

B. Re-ranking for Text Summarization

Although re-ranking work is relatively rare in the medical domain, it has received increasing attention from researchers working on general text summarization. Some studies have attempted to integrate contrastive learning into the re-ranking framework to enhance text summarization. For example, Zhong et al. adopted an extract-then-match framework for extractive text summarization, where candidate summaries were first extracted from a document and then re-ranked using contrastive learning based on their similarity with the document [24]. Based on this, Liu et al. proposed a two-stage framework on generative text summarization, where BART was used to generate candidate summaries and contrastive learning was employed to re-rank them [19]. Both studies utilized text semantic similarity to evaluate and re-rank the summaries. To more accurately evaluate the quality of candidate summaries, they further introduced a novel training paradigm that assumes a non-deterministic distribution assigning probability masses to different candidate summaries according to their quality [25]. Some studies have focused on combining multi-task learning with the re-ranking framework for text summarization. For instance, Ravaut et al. utilized a multi-task mixture-of-experts re-ranking framework to improve the performance of abstractive summarization systems, which transformed the assessment of candidate summary quality into a binary classification problem [12]. All the aforementioned work mentioned above is supervised, but some works also pay attention to the unsupervised methods. For example, Ravaut et al. developed a novel multi-objective re-ranker that aggregates the features from the candidate summary and source documents. This approach does not rely on any supervision information or require training neural models [26]. Additionally, recent research has been conducted on specific domains, apart from

the general domain. For instance, Elaraby et al. analyzed the argumentation present in legal documents. They utilized argument role information to generate candidate summaries, subsequently re-ranking them based on their alignment with the argument structure of the legal document [27].

While these approaches have achieved impressive results on open-domain datasets such as CNNDM and XSum [28], [29], their performance on the MQS task has not been evaluated. In this paper, we aim to thoroughly investigate by implementing two popular and robust models (SimCLS [19] and BRIO [25]) to test their performance on MQS. Additionally, we propose two novel re-rankers for the two-stage re-ranking framework, which significantly outperform the existing methods on MQS.

III. Methodology

A. The Proposed Framework

The framework of the proposed two-stage model, along with two re-rankers, is illustrated in Fig.1. The framework consists of two stages: one for generating multiple candidate summaries and another for re-ranking them to select the best one as the final summary. In the first stage, a Seq2Seq model is trained to generate several candidate summaries for medical questions. In the second stage, the candidate summaries are scored and re-ranked using a reference-free approach. Two re-rankers are proposed for the second stage. The first one is SimGate, which utilizes text semantic similarity and a confidence gate to select the final summary from the candidate ones. The second re-ranker is LLM, which utilizes a prompt template to guide a large language model to select the best summary for the original questions.

B. Candidate Summary Generation

1) Train a Seq2Seq Model: To generate candidate summaries, pretrained language model with Seq2Seq architecture, such as ProphetNet [9], PEGASUS [10], and BART [11] need to be finetuned. Among these models, BART has demonstrated superior performance on MQS task [7], [12], which is why we have chosen it as our base model. BART is trained using the maximum likelihood estimation (MLE) algorithm. For the i-th training sample $\{Q, S\}$, where Q, S denote the consumer health question and gold reference summary, respectively, MLE is equivalent to minimizing the sum of the negative likelihood of the l tokens $\{s_1, \cdots, s_j, ..., s_l\}$ in the reference summary S, i.e., to optimize the cross-entropy loss:

$$\mathcal{L}_{ce} = -\sum_{j=1}^{l}\sum_{s^*} p_{true}(s^*|Q, S_{<j}) \log p_{f_\theta}(s^*|Q, S_{<j}; \theta), \quad (1)$$

where s^* represents the token currently generated by the model. $S_{<j}$ refers to the partial reference sequence $\{s_0, \cdots, s_{j-1}\}$ and s_0 is a pre-defined start token. p_{true} denotes the one-hot distribution in the standard MLE framework. θ refers to the parameters of f and p_{f_θ} is the probability distribution entailed by these parameters.

2) Generate Candidate Summary: For a specific sample $\{Q, S\}$, the fine-tuned BART model $g(\cdot)$ can generate the probability distribution $\boldsymbol{D}$ for Q:

$$\boldsymbol{D} = g(Q). \quad (2)$$

Candidate summaries can be obtained using sampling algorithms, such as beam search or diverse beam search. In this study, beam search is employed to generate a set of candidate summaries $\mathbb{C} = \{C_1, \cdots, C_n\}$ for the given medical question Q:

$$\mathbb{C} = BeamSearch(\boldsymbol{D}, n), \quad (3)$$

where n represents the number of candidate summaries.

Following the retrieval of candidate summaries, we assess each group using evaluation metrics commonly employed in the MQS task, such as ROUGE [30], BERTScore [31], and BARTScore [32]. For scoring the candidate summaries, we use ROUGE in this study. The scored candidate summaries are represented as

$$\mathbb{C}^* = \{C_1^*, \cdots, C_i^*, \cdots, C_n^*\}, \quad (4)$$

where $C_1^*, \cdots, C_n^*$ are sorted in descending order based on their score. These scores are employed as supervision signals for model optimization, specifically during training rather than inference.

C. SimGate as Re-ranker

1) Training: We utilize the pre-trained language model RoBERTa [33] to encode the samples and obtain their embedding representations:

$$\begin{aligned} \mathbf{E}_Q &= embedding(Q), \\ \mathbf{E}_S &= embedding(S), \\ \mathbf{E}_{C_i^*} &= embedding(C_i^*). \end{aligned} \quad (5)$$

Following SimCLS [19], we employ a ranking loss to optimize the parameters of RoBERTa [33]:

$$\mathcal{L}_{rank} = \sum_i \max(0, sim(\mathbf{E}_Q, \mathbf{E}_{C_i^*}) - sim(\mathbf{E}_Q, \mathbf{E}_S)) + \sum_i \sum_{j>i} \max(0, sim(\mathbf{E}_Q, \mathbf{E}_{C_j^*}) - sim(\mathbf{E}_Q, \mathbf{E}_{C_i^*}) + \lambda_{ij}), \quad (6)$$

where $\lambda_{ij} = (j-i) * \lambda$ is the corresponding margin following [24], and λ is a hyper-parameter. $sim(\cdot)$ is cosine similarity.

2) Inference: During the inference phase, we begin by encoding the original medical question Q and its candidate summaries $\mathbb{C} = \{C_1, \cdots, C_n\}$ using the trained RoBERTa. Then, we calculate the similarity between each candidate summary C_i and Q to obtain similarity scores. Subsequently, we identify the candidate summary $C_{\max}$ with the highest similarity score:

$$C_{\max} = \arg\max_{C_i}(sim(C_i, Q)), \quad (7)$$

where $sim(C_i, Q)$ refers to the cosine similarity calculated between the embeddings of C_i and Q.

Based on our empirical observation, when the similarity between $C_{\max}$ and Q is very close to that of C_1 and Q, the model tends to make mistakes. To address this issue, we introduce a confidence-like gating mechanism to mitigate this problem:

$$C_{final} = \begin{cases} C_{\max}, sim(C_{\max}, Q) - sim(C_1, Q) > \eta \\ C_1, sim(C_{\max}, Q) - sim(C_1, Q) \leqslant \eta \end{cases}, \quad (8)$$

where η is the confidence value of the model and C_1 is the original output of BART model. By employing the gating mechanism during the inference phase, our model can mitigate certain misjudgments. It is important to note that the gating mechanism is only utilized during the inference phase.

D. LLM as Re-ranker

Large language models (LLMs) have recently demonstrated remarkable comprehension abilities in various natural language processing (NLP) tasks. In this study, our goal is to investigate whether LLMs can effectively assess the quality of candidate summaries, thereby serving as reliable re-rankers. To the best of our knowledge, there is no previous work specifically focusing on the task of re-ranking in the context of MQS. Thus, we propose, for the first time, the use of an LLM as a re-ranker for the MQS task. To accomplish this, we formulate the re-ranking task of MQS candidate summaries as an answering task for multiple-choice questions. Furthermore, we design appropriate prompts that leverage the powerful comprehension ability of LLM, enabling us to effectively re-rank candidate summaries.

One crucial step in employing a few-shot LLM as a re-ranker is to construct a suitable prompt. Our objective is to create an appropriate prompt template consisting of a historical prompt, question, and options. The historical prompt is formed by gathering four samples from the validation set that uses the same prompt template, thereby creating a historical dialogue record. In each round of the dialogue, we provide the answer part. By employing the historical prompt, the large language model can comprehensively understand the task requirements and generate the output in the specified format aligned with the prompted answer options. The overall structure of the prompt template is as shown in Table II.

After obtaining the answer using the query constructed with the above prompt template, we employ a simple matching extraction algorithm to obtain the final candidate summary.

TABLE II
PROMPT TEMPLATE

Prompt Template
$[H]$ Answer the following question. Please provide only one answer option, no additional information is required. Question: Which option is the best summary of the following content? Content: $[Q]$ Option: $\{[A].[C_1]\}$ $\{[B].[C_2]\}$ ... $\{[Option].[C_n]\}$ Answer:

where H represents the historical prompt. Q is the original medical question and C_i denotes the i-th candidate summary in $\mathbb{C}$.

IV. EXPERIMENTS

A. Datasets

We conduct the experiments on four medical question summarization datasets, i.e., MeQSum, CHQ-Summ, iCliniq, and HealthCareMagic. **MeQSum**, created by Ben Abacha et al. [20], is an MQS dataset obtained from a collection distributed by the U.S. National Library of Medicine. Each question in the dataset is accompanied by expert-provided summaries. **CHQ-Summ**, developed by Yadav et al. [34], consists of data extracted from the Yahoo! Answers L6 corpus, which was manually annotated by six experts in medical informatics and the medical field. Initially, **iCliniq** and **HealthCareMagic** were derived from the MedDialog dataset by Mrini et al. [5], however, Wei et al. [7] identified data leakage issues in the original dataset and subsequently restructured these two datasets. We adopt the modified and fair version of these two datasets. Table III displays the statistics of the four datasets.

TABLE III
DATASETS STATISTICS

Datasets	Examples			Avg. Words	
	Train	Valid	Test	CHQ	FAQ
MeQSum	400	100	500	70	12
CHQ-Summ	800	300	407	176	13
iCliniq	16,556	2,069	2,071	114	13
HealthCareMagic	180,697	22,587	22,588	93	11

B. Experimental Settings

We utilize BART-large [11] from HuggingFace as the Seq2Seq component in our model. The learning rate is set to 1e-5, and the batch size is set to 16. In the Adam optimizer, the values for β_1 and β_2 are set to 0.9 and 0.999, respectively.

For SimGate re-ranker, we utilize RoBERTa-base [33] from HuggingFace as the encoder. The maximum learning rate is set to 2e-3, and warmup_steps is set to 10000. The batch size is 16, and the number of candidate summaries is set to 16 (same as the previous works [19], [25]). For the value of λ in Eq. (6), it is set to 0.01 for MeQSum and CHQ-Summ, and 0.001 for iCliniq and HealthCareMagic. The values for η in Eq. (8) are set to 0.1, 0.1, 0.01, and 0.005 for MeQSum, CHQ-Summ, iCliniq, and HealthCareMagic, respectively.

Regarding the large language models for LLM re-ranker, we locally deploy three open-source models: LLaMA-3.1-8B-Instruct [35], GLM-4-9b-chat [36], and Qwen2.5-14B-Instruct [37]. The number of candidate summaries is set to 4. The parameter do_sample is set to False for stable results.

TABLE IV

THE RESULTS MARKED WITH * ARE TAKEN FROM OTHER PAPERS, WHILE ALL OTHER RESULTS ARE OBTAINED BY RUNNING AND FINETUNING THE CORRESPONDING CODE. *BART* IS ITALIC BECAUSE IT IS EMPLOYED AS THE FUNDAMENTAL MODEL IN THE FIRST STAGE FOR ALL TWO-STAGE SYSTEMS. THE BEST PERFORMANCE ON EACH METRIC IS BOLD. IMPROVEMENT IS COMPUTED BY COMPARING THE BEST PERFORMANCE ACHIEVED BY OUR METHODS AND THE PERFORMANCE OF BART. FOR LARGE LANGUAGE MODELS, WHETHER APPLIED FOR SUMMARIZATION OR RE-RANKING, THE OFFICIAL ORIGINAL WEIGHTS ARE USED.

Category	Model	MeQSum			CHQ-Summ			iCliniq			HealthCareMagic		
		R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
One Stage Models	LLaMA-3.1-8B-Instruct [35]	41.14	19.38	36.19	32.60	11.89	28.54	31.30	11.43	25.78	29.63	8.81	24.43
	GLM-4-9B-Chat [36]	43.70	20.31	39.42	36.07	14.66	32.33	33.14	11.88	27.17	32.27	10.03	27.37
	Qwen2.5-14B-Instruct [37]	42.11	19.73	37.60	35.06	14.68	30.94	34.30	13.16	27.97	30.20	9.80	25.04
	T5 [38]	34.47	17.18	30.54	35.17	18.87	32.33	39.25	20.84	34.92	38.09	18.41	34.99
	PEGASUS [10]	43.18	26.15	40.87	35.89	18.86	33.27	36.09	18.30	31.45	35.17	15.75	30.15
	ProphetNet [9]	44.40	26.93	41.57	40.46	22.80	38.13	39.13	20.15	34.01	30.34	12.00	26.00
	ProphetNet + QTR + QFR [18]*	45.52	27.54	48.19	-	-	-	-	-	-	-	-	-
	Joint Learning with DA [5]*	48.50	29.70	44.90	-	-	-	-	-	-	-	-	-
	RQE + MTL + DA [6]*	49.20	29.50	44.80	-	-	-	-	-	-	-	-	-
	QFCL [2]*	51.48	34.16	49.08	42.18	23.48	39.81	40.93	22.07	36.27	43.36	23.39	40.44
	BART [11]	*51.79*	*34.94*	*49.16*	*42.40*	*24.00*	*39.79*	*40.66*	*21.87*	*35.83*	*43.64*	*23.60*	*40.55*
	ECL [7]*	52.85	**36.06**	50.48	43.16	24.26	40.46	41.31	22.27	36.68	43.52	23.75	40.56
Two Stage Models	BRIO-Ctr [25]	49.77	30.68	47.24	39.49	19.98	36.97	39.18	20.10	34.97	41.10	20.57	38.16
	SimCLS [19]	52.41	33.88	49.90	42.46	23.06	39.58	41.85	22.18	36.97	44.58	23.21	41.24
	Our Methods												
	BART + LLaMA-3.1-8B-Instruct	50.84	32.26	47.41	43.59	25.20	40.94	40.87	22.19	36.32	43.56	22.51	40.02
	BART + GLM-4-9b-chat	51.72	32.90	48.29	43.31	24.83	40.47	41.73	22.81	36.93	43.80	22.87	40.25
	BART + Qwen2.5-14B-Instruct	52.58	33.48	49.17	43.20	24.66	40.40	41.41	22.59	36.66	43.80	22.71	40.22
	BART + SimGate re-ranker	**53.33**	35.83	**51.17**	**44.00**	**25.47**	**41.15**	**41.94**	**22.80**	**37.13**	**44.96**	**23.92**	**41.66**
	Improvement (vs. BART)	↑ 2.97%	↑ 2.55%	↑ 4.09%	↑ 3.77%	↑ 6.13%	↑ 3.42%	↑ 3.15%	↑ 4.25%	↑ 3.63%	↑ 3.02%	↑ 1.36%	↑ 2.74%

We adopt ROUGE [30] as the evaluation metric. R1, R2, and RL metrics denote ROUGH-1, ROUGH-2, and ROUGH-L, respectively. All experiments are conducted with one NVIDIA A100 40GB GPU.

C. Experimental Results

To demonstrate the effectiveness of our proposed framework equipped with SimGate and LLM re-ranker, we compare them with the popular and state-of-the-art baselines. The overall results are shown in Table IV, from which we have the following observations:

First, the upper panel of the table shows the performance of traditional one-stage models, including the state-of-the-art ECL model. Although these one-stage models show good performance, their overall performance metrics are inferior to those of two-stage models in the low panel of the table. All of the one-stage models utilize beam search to find one unique summary, failing to fully exploit the vast search space. Besides, they neglect the mismatching between the learning objective of Seq2Seq and the evaluation metrics of MQS. This leads to their inferiority compared to the two-stage systems.

Second, although both BRIO-Ctr and SimCLS are two-stage models, their performance on MQS datasets is still unsatisfactory. For BRIO-Ctr, its performance is worse than BART, whose inferiority may be caused by the significant disparity in these datasets. As shown in Table III, the average length of FAQ in MQS ranges from 11 to 13 words. However, the average length of FAQ in the datasets adopted by BRIO-Ctr ranges from 23 to 123 words. The brevity of the FAQ data might limit BRIO's ability to learn the allocation probabilities when trained on the MQS dataset, leading to inferior performance. SimCLS effectively improves the performance of BART on R1 and RL metrics, with a decrease in the R2 metric. This decline might be due to the fact that SimCLS occasionally misjudge some samples with difficult-to-distinguish similarities.

Third, comparing with the fundamental BART model, our proposed SimGate re-ranker demonstrates the superiority in all metrics across four datasets. Besides, except for the metric R2 on MeQSum, SimGate re-ranker outperforms all baselines, achieving a new state-of-the-art result. The success is attributed to the re-ranking strategy, which combines text semantic similarity and a confidence gate to select the final summary. Especially, the confidence gate is simple yet very effective to reduce mistakes.

Fourth, compared to the basic BART model, our proposed LLM re-ranker with the few-shot setting performs well on the MeQSum, CHQ-Summ, iCliniq, and HealthCareMagic datasets. The overall experimental results indicate that large language models have a certain level of understanding in selecting candidate summaries, making them a promising re-ranker.

Finally, comparing the best performance achieved by our proposed re-rankers and that of the fundamental BART model, the improvement is significant, ranging from 1.36% to 6.13%. The great improvement demonstrates the superiority and effectiveness of our proposed re-rankers on MQS.

D. Effect of η in SimGate Re-ranker

For SimGate re-ranker, the threshold value η of the confidence gate is of utmost importance. According to our empirical observation, it exhibits a strong correlation with the value of λ in $\mathcal{L}_{rank}$ in Eq. (6). We have observed that when the value of λ decreases, the semantic similarity of the encoded representations becomes concentrated, while conversely, when the value of λ increases, the similarity becomes dispersed. Based on this observation, we propose an empirical approach to determine the value of η:

$$\eta = [\lambda \times 10/2, \lambda \times 10]. \tag{9}$$

TABLE V
INFLUENCE OF GATE THRESHOLD η IN SIMGATE RE-RANKER. ALL FOUR DATASETS EMPLOY 16 CANDIDATE SUMMARIES. FOR THE MEQSUM AND CHQ-SUMM DATASETS, THE VALUE OF λ IN $\mathcal{L}_{rank}$ IS SET TO 0.01, WHILE FOR THE OTHER TWO DATASETS, IT IS SET TO 0.001.

Dataset	η	R1	R2	RL
MeQSum	0	52.41	33.88	49.90
	0.001	52.39	33.88	49.88
	0.005	52.74	34.26	50.26
	0.01	52.80	34.31	50.35
	0.05	53.50	35.57	51.02
	0.1	**53.33**	**35.83**	**51.17**
CHQ-Summ	0	42.46	23.06	39.58
	0.001	42.44	23.06	39.56
	0.005	42.37	22.98	39.57
	0.01	42.37	23.16	39.55
	0.05	43.13	24.51	40.37
	0.1	**44.00**	**25.47**	**41.15**
iCliniq	0	41.85	22.18	36.97
	0.001	41.91	22.30	37.06
	0.005	41.94	22.59	37.09
	0.01	**41.94**	**22.80**	**37.13**
	0.05	40.66	21.87	35.83
	0.1	40.66	21.87	35.83
HealthCareMagic	0	44.58	23.21	41.24
	0.001	44.71	23.41	41.38
	0.005	**44.96**	**23.92**	**41.66**
	0.01	44.69	23.90	41.40
	0.05	43.14	23.15	40.08
	0.1	43.14	23.15	40.08

As shown in Table V, we validate the performance of η using this empirical approach. According to the results, for MeQSum and CHQ-Summ, we set λ to 0.01 and η to 0.1. It is evident that both of the parameter values effectively improve the model's performance. Similarly, for iCliniq and HealthCareMagic, we set λ to 0.001, and set η to 0.01 and 0.005 respectively.

E. Comparison of Re-ranking Selection Accuracy Between SimCLS and SimGate

Both SimClS and SimGate are two-stage systems based on BART, which employ a similarity-based re-ranking mechanism to select the best summary from candidate ones. For a sample, if the ROUGE-L score of a summary selected by a re-ranker in the second stage is greater than or equal to the score of the original summary outputted by BART in the first stage, the sample's summary is chosen rightly by the re-ranker. If not, the summary is chosen wrongly by the re-ranker.

To further investigate the effectiveness of re-ranking mechanism of SimCLS and SimGate, we define the re-ranking selection accuracy (RSA), which measures how many samples' summary can be chosen rightly in the second stage. We compare RSA between SimCLS and SimGate. As shown in Fig. 2, SimGate re-ranker demonstrates a significant and consistent superiority over SimCLS across four datasets, which verifies the significance of confidence gate in SimGate re-ranker.

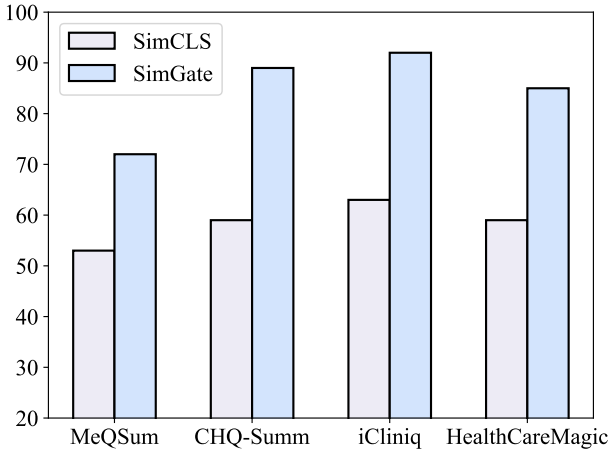

Fig. 2. Comparison of Re-ranking Selection Accuracy between SimCLS and SimGate

V. CONCLUSION

In this paper, to solve the mismatching problem of the objective function of Seq2Seq and evaluation metrics of MQS, we propose to improve MQS through re-ranking. We introduce a two-stage framework into the MQS task, where the first stage employs a Seq2Seq model to generate candidate summaries, and the second one utilizes a re-ranker to evaluate the candidates and select the optimal one. Moreover, we propose two novel re-rankers, named SimGate and LLM re-rankers. The former evaluates candidate summaries based on text semantic similarity and a confidence gate, while the latter leverages the powerful comprehension capability of large language models to select better summaries. The extensive experiments on four datasets validate the effectiveness of our proposed re-rankers. In the future, we plan to delve deeper into exploring the potential of large language models as re-rankers, aiming to better understand their capabilities and optimize their application in various ranking and retrieval tasks.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China (No.2022YFC3302101) and the National Natural Science Foundation of China (No.62376130), Shandong Provincial Natural Science Foundation (No.ZR2022MF243), Program of New Twenty Policies for Universities of Jinan (No.202333008), Program of Innovation Improvement of Shandong (No.2024TSGC0062, No.2023TSGC0182), Pilot Project for Integrated Innovation of Science, Education, and Industry of Qilu University of Technology (Shandong Academy of Sciences) (No.2024ZDZX08).

REFERENCES

[1] S. Wei, X. Peng, H. Guan, L. Geng, P. Jian, H. Wu, and W. Lu, "Multi-view contrastive learning for medical question summarization," in *Proceedings of the 27th International Conference on Computer Supported Cooperative Work in Design*, 2024, pp. 1826–1831.

[2] M. Zhang, S. Dou, Z. Wang, and Y. Wu, "Focus-driven contrastive learning for medical question summarization," in *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, 2022, pp. 6176–6186.

[3] C. Lei, V. Efthymiou, R. Geis, and F. Özcan, "Expanding query answers on medical knowledge bases," in *Proceedings of the 2020 International Conference on Extending Database Technology (EDBT)*, 2020, pp. 567–578.

[4] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–23, 2019.

[5] K. Mrini, F. Dernoncourt, W. Chang, E. Farcas, and N. Nakashole, "Joint summarization-entailment optimization for consumer health question understanding," in *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations (NLPMC)*, 2021, pp. 58–65.

[6] K. Mrini, F. Dernoncourt, S. Yoon, T. Bui, W. Chang, E. Farcas, and N. Nakashole, "A gradually soft multi-task and data-augmented approach to medical question understanding," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021, pp. 1505–1515.

[7] W. Lu, S. Wei, X. Peng, Y.-F. Wang, U. Naseem, and S. Wang, "Medical question summarization with entity-driven contrastive learning," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 4, pp. 1–19, 2024.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 6000–6010.

[9] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "ProphetNet: Predicting future n-gram for sequence-to-sequencepre-training," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 2401–2410.

[10] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 11 328–11 339.

[11] 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 (ACL)*, 2020, pp. 7871–7880.

[12] M. Ravaut, S. Joty, and N. Chen, "SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 4504–4524.

[13] M. Sänger, L. Weber, and U. Leser, "WBI at MEDIQA 2021: Summarizing consumer health questions with generative transformers," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 86–95.

[14] Y. He, M. Chen, and S. Huang, "damo_nlp at MEDIQA 2021: Knowledge-based preprocessing and coverage-oriented reranking for medical question summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 112–118.

[15] S. Balumuri, S. Bachina, and S. Kamath, "SB_NITK at MEDIQA 2021: Leveraging transfer learning for question summarization in medical domain," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 273–279.

[16] K. Mrini, F. Dernoncourt, S. Yoon, T. Bui, W. Chang, E. Farcas, and N. Nakashole, "UCSD-Adobe at MEDIQA 2021: Transfer learning and answer sentence selection for medical summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 257–262.

[17] S. Yadav, M. Sarrouti, and D. Gupta, "NLM at MEDIQA 2021: Transfer learning-based approaches for consumer question and multi-answer summarization," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 291–301.

[18] S. Yadav, D. Gupta, A. B. Abacha, and D. Demner-Fushman, "Reinforcement learning for abstractive question summarization with question-aware semantic rewards," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021, pp. 249–255.

[19] Y. Liu and P. Liu, "SimCLS: A simple framework for contrastive learning of abstractive summarization," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2021, pp. 1065–1072.

[20] A. B. Abacha and D. Demner-Fushman, "On the summarization of consumer health questions," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 2228–2234.

[21] A. B. Abacha, Y. M'rabet, Y. Zhang, C. Shivade, C. Langlotz, and D. Demner-Fushman, "Overview of the MEDIQA 2021 shared task on summarization in the medical domain," in *Proceedings of the 20th Workshop on Biomedical Language Processing (BioNLP)*, 2021, pp. 74–85.

[22] W. Lu, G. Zhang, X. Peng, H. Guan, and S. Wang, "Medical entity disambiguation with medical mention relation and fine-grained entity knowledge," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

[23] G. Zhang, X. Peng, T. Shen, G. Long, J. Si, L. Qin, and W. Lu, "Extractive medical entity disambiguation with memory mechanism and memorized entity information," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 13 811–13 822.

[24] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X.-J. Huang, "Extractive summarization as text matching," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 6197–6208.

[25] Y. Liu, P. Liu, D. Radev, and G. Neubig, "BRIO: Bringing order to abstractive summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, pp. 2890–2903.

[26] M. Ravaut, S. Joty, and N. Chen, "Unsupervised summarization re-ranking," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[27] M. Elaraby, Y. Zhong, and D. Litman, "Towards argument-aware abstractive summarization of long legal opinions with summary reranking," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[28] K. M. Hermann, T. Kočiskỳ, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, 2015, pp. 1693–1701.

[29] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! Topic-Aware convolutional neural networks for extreme summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 1797–1807.

[30] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proceedings of the 2019 International Conference on Learning Representations (ICLR)*, 2019.

[32] W. Yuan, G. Neubig, and P. Liu, "BARTScore: Evaluating generated text as text generation," in *Proceedings of the 34th Advances in Neural Information Processing Systems (NIPS)*, 2021, pp. 27 263–27 277.

[33] 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," *arXiv preprint arXiv:1907.11692*, 2019.

[34] S. Yadav, D. Gupta, and D. Demner-Fushman, "CHQ-Summ: A dataset for consumer healthcare question summarization," *arXiv preprint arXiv:2206.06581*, 2022.

[35] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The LLaMA 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[36] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao *et al.*, "ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools," *arXiv preprint arXiv:2406.12793*, 2024.

[37] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[38] 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," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.