Better Late Than Never: Model-Agnostic Hallucination Post-Processing Framework Towards Clinical Text Summarization

Songda Li[1], Yunqi Zhang[1], Chunyuan Deng[2], Yake Niu[1], Hui Zhao[1,3,*]
[1]Software Engineering Institute, East China Normal University
[2]Georgia Institute of Technology
[3]Shanghai Key Laboratory of Trustworthy Computing, Shanghai, China
{songda.li,yunqi.zhang,yake.niu}@stu.ecnu.edu.cn
cdeng73@gatech.edu
hzhao@sei.ecnu.edu.cn

Abstract

Clinical text summarization has proven successful in generating concise and coherent summaries. However, these summaries may include unintended text with hallucinations, which can mislead clinicians and patients. Existing methods for mitigating hallucinations can be categorized into task-specific and task-agnostic approaches. Task-specific methods lack versatility for real-world applicability. Meanwhile, task-agnostic methods are not model-agnostic, so they require retraining for different models, resulting in considerable computational costs. To address these challenges, we propose MEDAL, a model-agnostic framework designed to post-process medical hallucinations. MEDAL can seamlessly integrate with any medical summarization model, requiring no additional computational overhead. MEDAL comprises a medical infilling model and a hallucination correction model. The infilling model generates non-factual summaries with common errors to train the correction model. The correction model is incorporated with a self-examination mechanism to activate its cognitive capability. We conduct comprehensive experiments using 11 widely accepted metrics on 7 baseline models across 3 medical text summarization tasks. MEDAL demonstrates superior performance in correcting hallucinations when applied to summaries generated by pre-trained language models and large language models.[1]

1 Introduction

Given the widespread use of electronic health records (EHR) (e.g., health questions, radiology reports, and doctor-patient dialogues), medical text summarization enhances healthcare efficiency (Van Veen et al., 2023b). However, manual summarization is notably laborious and time-consuming (Rink et al., 2023), leading to potential errors under overwhelming workloads. Encouragingly, there is a rising focus on clinical text summarization (He et al., 2021; Dai et al., 2021; Joshi et al., 2020) to alleviate clinicians' burden and enable them to refocus on patient-centered care.

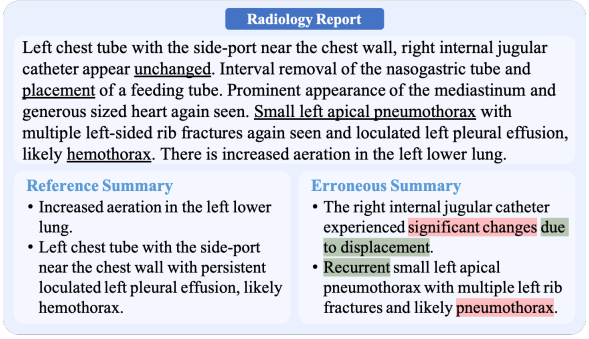

Figure 1: **A radiology report with its reference summary and an erroneous summary generated by a summarization model.** Intrinsic errors are highlighted in red, signifying sections that should remain "unchanged" and "hemothorax". Extrinsic errors are highlighted in green, indicating the issues of *imposing causality* and *introducing additional details.*

However, clinical text summarization may engender hallucinations, resulting in intrinsic hallucinations (facts that contradict the source text) or extrinsic hallucinations (facts that cannot be directly inferred from the source text) (Maynez et al., 2020; Zhang et al., 2023), as illustrated in Figure 1. Hallucinations may mislead clinicians and patients, posing significant medical risks. Mitigating hallucinations is imperative for the seamless integration of automated summarization into clinical workflows.

Numerous methods have been proposed to address hallucinations in general summarization tasks (Ji et al., 2023; Li et al., 2022). However, these methods lack tailored designs for medical contexts, limiting their direct application to the medical domain. Recently, specific techniques for tackling hallucinations in medical summarization have been introduced. Adams et al. (2022) suggest selectively

*Corresponding author.
[1]Our code is available at https://github.com/lisdarr/MEDAL.

Findings of the Association for Computational Linguistics ACL 2024, pages 995–1011

rewrite unsupported reference sentences to improve the training set. Meanwhile, Zhang et al. (2020) and Alambo et al. (2022) focus on enhancing summarization models via reinforcement learning and knowledge-guided multi-objective optimization, respectively. Nevertheless, these methods are unsuitable for clinical deployment as they are confined to a specific type of medical text. To address this challenge, Zhang et al. (2023) introduce a fine-tuning strategy with wide applicability. However, its contrastive learning component demands high-quality training data and substantial computational costs. Moreover, the rise of large language models (LLMs) escalates fine-tuning costs due to the increasing model parameters.

We propose MEDAL, a model-agnostic post-processing framework designed to address medical hallucinations in various summarization tasks. MEDAL comprises two components: 1) a *hallucination correction model* for post-processing, and 2) a *medical infilling model* to generate training data for the correction model. The *hallucination correction model* is built upon Flan-T5 (Chung et al., 2022), a model fine-tuned with instructions. Our hallucination correction model fully leverages the cognitive capabilities of Flan-T5 through self-examination. The model is guided to examine hallucinations in a model-generated summary and then correct the summary based on its self-examination. Specifically, we introduce a `self-examination` token that triggers the hallucination correction model to refine its generation process. Furthermore, we employ the *medical infilling model* to create non-factual summaries with errors aligned with those found in model-generated summaries. This approach aims to enhance the correction model's capacity to rectify common errors in automated medical summaries. Specifically, we enhance a traditional infilling model (Donahue et al., 2020) by constructing medical Cloze questions. These medical Cloze questions help incorporate medical knowledge into the infilling model during training and are utilized to generate non-factual summaries during inference.

We conduct extensive experiments to evaluate the performance of MEDAL across three tasks: health question summarization, radiology report summarization, and doctor-patient dialogue summarization. First, MEDAL has consistently achieved optimal or near-optimal performance using widely accepted general quality metrics and faithfulness metrics across three medical text summarization tasks. Second, MEDAL effectively rectifies hallucinations in summaries generated by LLMs, demonstrating its potential for future application. Third, we demonstrate MEDAL's robustness with the imbalanced training set. Finally, further studies are conducted to comprehensively explain the effectiveness of self-examination.

2 Related Work

Medical text summarization Medical text summarization encompasses three key tasks: 1) *health question summarization* condenses an inquiry into a concise question to elicit accurate responses; 2) *radiology report summarization* condenses a detailed "findings" section into an "impression" section, capturing the most pertinent and actionable information; and 3) *doctor-patient dialogue summarization* aims to encapsulate full conversations into a succinct paragraph.

Health question summarization is introduced by Ben Abacha and Demner-Fushman (2019). Recent models have achieved performance improvements using reinforcement learning (Yadav et al., 2021a), contrastive learning (Zhang et al., 2022), transfer learning (Yadav et al., 2021b), multi-task learning (Mrini et al., 2021), and question-aware transformer models (Yadav et al., 2022). For *radiology report summarization*, most prior works leverage pre-extracted knowledge like ontologies and word graphs (MacAvaney et al., 2019; Sotudeh Gharebagh et al., 2020; Hu et al., 2021). Hu et al. (2022b) and Karn et al. (2022) advocate for contrastive learning and reinforcement learning, respectively. Recently, attention has shifted towards employing LLMs (Van Veen et al., 2023a; Karn et al., 2023). *Doctor-patient dialogue summarization* is early studied with pointer generator networks (Joshi et al., 2020). Krishna et al. (2021) explore summarization methods, while Navarro et al. (2022) focus on the few-shot setting. Zhang et al. (2021) focus on handling long conversations by fine-tuning BART. MedDialog (Zeng et al., 2020) and ACI-BENCH (Yim et al., 2023) introduce datasets to advance the research.

Differing from the task-specific methods mentioned above, we propose a model-agnostic framework suitable for all these tasks, showcasing consistent strong performance across these tasks.

Mitigating hallucinations in medical text summarization Researches on hallucinations in medical text summarization primarily focus on training

data and training method optimization. Adams et al. (2022) selectively rewrite unsupported reference sentences to better reflect source data. Zhang et al. (2020) utilize reinforcement learning, while Xie et al. (2023) propose FactReranker which selects the best summary based on an estimated factual consistency score. Krishna et al. (2021) propose an algorithm that extracts and clusters pertinent utterances to generate one summary sentence per cluster. Zhang et al. (2023) introduce a fine-tuning strategy using contrastive learning. However, its resource-intensive nature and requirement for high-quality training data pose challenges for application.

To the best of our knowledge, we are the first to propose a post-processing framework to rectify summaries generated by medical summarization models. Moreover, our framework is versatile and lightweight compared to previous methods.

3 Method

In this section, we first provide an overview of our proposed framework MEDAL (§3.1). Sequentially, we detail the construction of medical Cloze questions (§3.2). Finally, we delve into the architectures of the medical infilling model (§3.3) and the hallucination correction model (§3.4).

3.1 Overview

For the medical text summarization task, the summarization model generates a summary s' based on a given medical source text m. However, the model-generated summary s', namely a dafted summary, may contain hallucinations. Our objective is to rectify hallucinations and align the corrected summary s more closely with the reference summary r.

MEDAL is a model-agnostic hallucination post-processing framework without retraining summarization models, resulting in low costs and wide application. As illustrated in Figure 2, the MEDAL framework involves three essential steps.

First, we mask terms that are error-prone during generation to construct medical Cloze questions. These masked terms are selected based on the taxonomy of faithfulness errors in medical summaries (Zhang et al., 2023).

Second, the medical infilling model, denoted as $\mathcal{M}_I$, utilizes the medical Cloze questions to generate non-factual summaries. During the training phase, the model's primary objective is to acquire medical knowledge. Consequently, it is trained to complete the medical Cloze task. During the inference phase, motivated by Goyal and Durrett (2020), we select lower-ranked token candidates as output to create non-factual summaries. These non-factual summaries manifest subtle deviations from the reference summaries.

Third, the hallucination correction model, denoted as $\mathcal{M}_C$, is built upon Flan-T5 (Chung et al., 2022). To activate the cognitive capability of Flan-T5, we guide it towards a self-examination process. This innovative addition allows the model to initially examine hallucinations in the summary. Subsequently, the model refines its generation process. $\mathcal{M}_C$ is trained using both the non-factual summaries generated by $\mathcal{M}_I$ and reference summaries, thereby enhancing its capacity to rectify hallucinations in summaries.

3.2 Construction of medical Cloze questions

We employ diverse tools to identify error-prone terms in a medical text. Common errors appear in *entity*, *entity relation*, *quantifier*, and *negation*. We outline them in Appendix A. We utilize MedCAT[2] to extract medical entities. Additionally, we leverage Stanford OpenIE[3] to detect the relation triples. Quantifiers are identified through regular expression matching techniques, while negations are recognized based on a predefined set of negation terms {"no", "nope", "doesn't", "don't", "not"}.

Given a medical text p, we use the mentioned tools to identify all error-prone terms in the text, forming the set T. Subsequently, each identified error-prone term is replaced with a `[MASK]` token to form a Cloze question. This process is formalized as the function $\textsc{GenCloze}(p, T)$, which yields the set of medical Cloze questions, denoted as Q:

$$\begin{aligned} Q =& \textsc{GenCloze}(p, T) \\ =& \{(clz, t) \mid \\ & clz = \text{replace}(p, t, \texttt{[MASK]}), t \in T\} \end{aligned} \tag{1}$$

where the funtion $\text{replace}(p, t, \texttt{[MASK]})$ replaces the term t in the medical text p with "`[MASK]`" and returns the modified text.

[2]MedCAT is an open-source medical concept annotation toolkit proposed by Kraljevic et al. (2021).

[3]Stanford OpenIE is an open information extraction tool proposed by Angeli et al. (2015). Open information extraction (OpenIE) aims to extract relation tuples from text without predefined relations. The relation name is just the text linking two arguments.

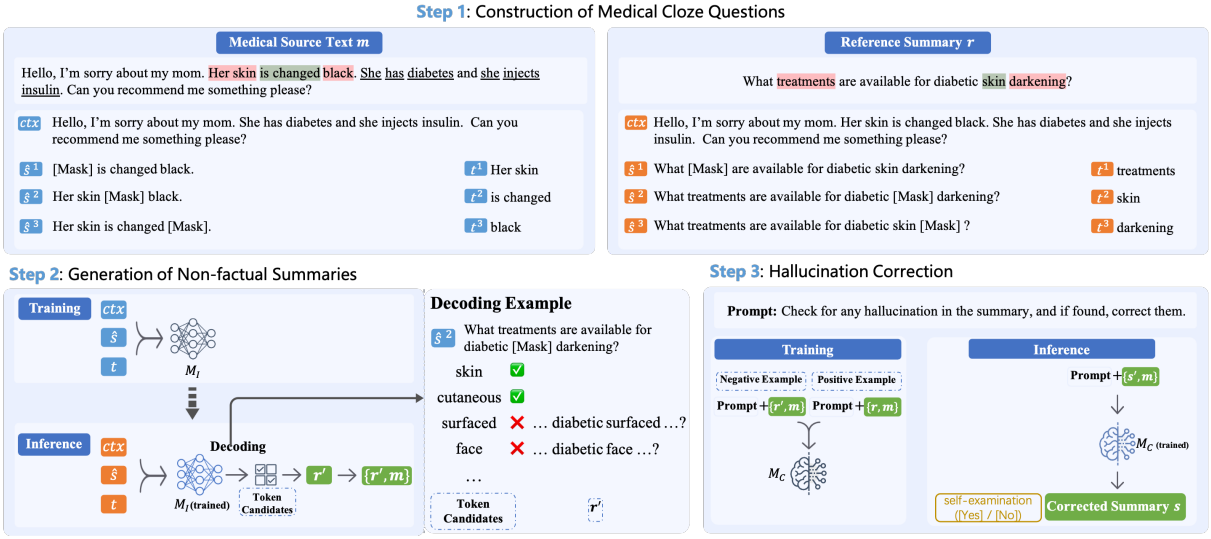

Figure 2: **The overall framework of MEDAL.** In Step 1, we construct medical Cloze questions derived from medical source texts and reference summaries. ctx denotes the context. $\hat{s}$ is the medical Cloze question and t represents its masked term. Notably, we only take one sentence in a medical source text as an example and other underlined terms also need to be masked. In Step 2, the infilling model $\mathcal{M}_I$ is trained using the dataset generated from medical source texts. During the inference phase, $\mathcal{M}_I$ predicts several token candidates, and among them, we select the lower-ranked token candidates to generate the non-factual summaries. In Step 3, the hallucination correction model $\mathcal{M}_C$ is trained using both the non-factual summaries and the reference summaries. The trained $\mathcal{M}_C$ is utilized to rectify hallucinations in model-generated summaries.

3.3 Medical infilling model

The medical infilling model $\mathcal{M}_I$ is based on BioBART (Yuan et al., 2022). The input is the concatenation of the sentence with the `[MASK]` token $\hat{s}$ and its context c, while the target output is the original error-prone term t replaced by the `[MASK]` token.

Training The training set is generated from medical source texts, where error-prone terms are identified and replaced with the `[MASK]` token. Algorithm 1 shows the method for generating the training set. During training, the model learns to predict the masked terms based on their context. By optimizing this objective, the model acquires a deeper understanding of medical terminology and context.

Inference The inference phase aims to produce non-factual summaries that manifest subtle deviations from the reference summaries. Therefore, we generate the input dataset utilizing the reference summaries. The detailed procedure is outlined in Algorithm 2.

During inference, the trained medical infilling model $\mathcal{M}_I$ employs beam-search decoding to generate various token candidates. An example is depicted in the Decoding Example section of Step 2 in Figure 2. Lower-ranked candidates often exhibit semantics or parts of speech similar to the original term but lack factual accuracy (Zhang et al., 2023). Therefore, we select lower-ranked candidates as non-factual replacements for the masked term. The method for selecting the lower-ranking range is detailed in Appendix B. The non-factual summary, denoted as r', is produced by replacing the `[MASK]` token in $\hat{s}$ with the token candidate.

3.4 Hallucination correction model

The backbone of the hallucination correction model $\mathcal{M}_C$ is Flan-T5 (Chung et al., 2022), known for its abilities in reasoning and generalization. To engage the cognitive capabilities of Flan-T5, we design an instruction (shown in Step 3 of Figure 2), directing the model to examine hallucinations before generating the corrected summary. We introduce a `self-examination` token, where `[YES]` indicates the presence of hallucinations and `[NO]` indicates their absence.

The input of the hallucination correction model comprises a prompt sentence, a medical source text m, and a summary, while the target output is a combination of either `[NO]` or `[YES]` along with the corrected summary s. The detailed input and output formats are listed in Appendix D.5.

Training To prevent the model from modifying factual summaries, we structure the training set

Algorithm 1 Training Set Generation

Input: The medical source text, m; the set of error-prone terms identified in the medical source text, T_m;
Output: The training set, D_{train};

$Q \leftarrow$ GENCLOZE(m, T_m)
$D_{train} \leftarrow$ empty set
for each $(cloze, term)$ in Q **do**
    $c \leftarrow$ ""
    $sents \leftarrow$ sentence_tokenize($cloze$)
    for each $sent$ in $sents$ **do**
        if "[MASK]" in $sent$ **then**
            $\hat{s} \leftarrow sent$
            $t \leftarrow term$
        else
            $c = c + sent$
        end if
    end for
    add $(\hat{s}, c, t)$ to set D_{train}
end for
return D_{train}

Algorithm 2 Inference Dataset Generation

Input: The medical source text, m; the reference summary, r; the set of error-prone terms identified in the reference summary, T_r;
Output: The inference set, D_{infer};

$Q \leftarrow$ GENCLOZE(r, T_r)
$D_{infer} \leftarrow$ empty set
for each $(cloze, term)$ in Q **do**
    $\hat{s} \leftarrow cloze$
    $c \leftarrow m$
    $t \leftarrow term$
    add $(\hat{s}, c, t)$ to set D_{infer}
end for
return D_{infer}

with both positive samples and negative samples. Positive samples account for 40% of the dataset, while negative samples consist of the remaining 60%. The summaries in positive samples are from reference summaries, whereas the summaries in negative samples are non-factual summaries generated by $\mathcal{M}_I$.

Inference During inference, the summary of each input is the model-generated summary s'. We extract the corrected summaries from the output of the correction model.

4 Experiment Setup

4.1 Datasets

To illustrate MEDAL's efficacy in medical text summarization, we conduct experiments on three distinct types of medical texts: health questions, radiology reports, and doctor-patient dialogues. The statistics of the datasets are listed in Table 1.

Health questions We employ Health Question Summarization (HQS) from MEDIQA 2021 shared task 1 (Ben Abacha et al., 2021). The training set is from the MeQSum Dataset (Ben Abacha and Demner-Fushman, 2019). The validation and test sets include diverse consumer health questions received by the U.S. National Library of Medicine, along with expert-generated gold summaries.

Radiology reports We utilize Radiology Report Summarization (RRS) from MEDIQA 2021 shared task 3 (Ben Abacha et al., 2021). The source of RRS is the MIMIC-CXR chest X-ray report dataset[4] and the Indiana University chest X-ray dataset[5]. The validation set is selected from the Indiana dataset, and the test set combines the Indiana dataset with newly released reports drawn from the Stanford Health Care system.

Doctor-patient dialogues Most datasets are private due to ethical limitations. Therefore, we utilize the public ACI-BENCH dataset (Yim et al., 2023) with full doctor-patient conversations and associated clinical notes.[6] Van Veen et al. (2023b) select 126 samples containing an "assessment and plan" section from the dataset. We follow their work and divide the 126 samples according to the proportions of each set in the original ACI-BENCH dataset.

4.2 Baselines

Pre-trained language models We fine-tune **Pegasus-large** (Zhang et al., 2020), a abstractive summarization model; and **BioBART-large** (Yuan et al., 2022), a biomedical generative model.

Task-agnostic medical summarization model We evaluate MEDAL against the state-of-the-art (SOTA) model **FAMESUMM** (Zhang et al., 2023), which performs contrastive learning to encourage the faithful generation of medical terms.

[4] https://physionet.org/content/mimic-cxr/2.0.0/
[5] https://openi.nlm.nih.gov/faq#collection
[6] A clinical note contains four sections: 1) subjective information reported by the patient; 2) objective observations; 3) assessments made by the doctor; and 4) a plan for future care (Krishna et al., 2021).

Task	Description	Dataset	Train	Dev	Test	Avg. Number of Tokens	
						Input	Target
Health Question Summarization	verbose → short question	HQS	1000	50	100	71	11
Radiology Report Summarization	findings → impression	RRS	91544	4000	600	52	14
Doctor-Patient Dialogue Summarization	dialogue → clinical note	ACI-BENCH	66	20	40	1512	211

Table 1: **The statistics of datasets.** We show the number of examples in the train/dev/test splits and the average number of tokens in each dataset.

Task-specific medical summarization models We select the SOTA methods for each medical summarization task. For *health question summarization*, we select **QFCL**[7] (Zhang et al., 2022) which employs contrastive learning. For *radiology report summarization*, we choose **WGSum+CL** (Hu et al., 2022b) combining graphs and contrastive learning. For *doctor-patient dialogue summarization*, task-specific methods are fine-tuned on large private datasets. Given the paucity of public training data, these methods are not suitable as baselines.

Large language models We select two open-source large language models: 1) **Llama-2 (7B)** (Touvron et al., 2023) and 2) **Med-Alpaca** (Han et al., 2023) specifically fine-tuned for medical question-answering and dialogue applications.

4.3 Metrics

We adopt comprehensive evaluation metrics.[8] In addition to the general quality metrics and the faithfulness metrics, we incorporate the latest evaluation metrics for automated medical note generation proposed by Ben Abacha et al. (2023).

General quality metrics Two reference-based metrics, ROUGE-1/2/L and BERTSCORE, are used to evaluate the general quality of summaries. ROUGE-1/2/L (Lin, 2004) is used to measure the count of overlapping units, while BERTSCORE (Zhang* et al., 2020) calculates a similarity score between the generated and reference summary.

General faithfulness metrics QUESTEVAL (Scialom et al., 2021) evaluates whether the generated and reference summaries contain the same information based on question answering. SUMMAC (Laban et al., 2022) measures factual consistency by natural language inference (NLI).

Task-specific faithfulness metrics FaR (Shing et al., 2021) and Concept F1 (C F1) (Joshi et al., 2020) are computed based on medical entities. Recently, Ben Abacha et al. (2023) introduce three metrics, MedBERTScore, MedBARTScore, and ClinicalBLEURT.

4.4 Implementation details

We use `biobart-large` (Yuan et al., 2022) and `flan-t5-base` (Chung et al., 2022) as the backbones of the infilling model $\mathcal{M}_I$ and the correction model $\mathcal{M}_C$, respectively. Notably, the length of the doctor-patient dialogues exceeds Flan-T5's maximum context length. We select relevant supporting sentences as the source input of the correction model. More details are presented in Appendix D.

5 Results and Analysis

In this section, we aim to answer the following research questions:

- **RQ1:** Does MEDAL perform well across different medical text summarization tasks on general quality metrics and faithfulness metrics (§5.1)?
- **RQ2:** Does the self-examination process activate MEDAL's cognitive capability (§5.2), and if affirmative, what contributes to its efficacy (§5.4)?
- **RQ3:** Does MEDAL exhibit robustness when trained with imbalanced samples (§5.3)?

5.1 Overall results

The overall results for the three tasks are shown in Table 2, 3, and 4, respectively.

Improvements over medical summarization models First, MEDAL achieves SOTA or near-SOTA performance in three medical text summarization tasks. Compared with the competitive baseline FAMESUMM, MEDAL requires fewer computational costs. FAMESUMM is a fine-tuning method which needs retraining when applied to different summarization models. Besides, FAMESUMM involves learning from contrastive summaries, leading to high requirements on GPU memory. In con-

[7] QFCL is a competitive method for health question summarization. **ProphetNet+QTR+QFR** (Yadav et al., 2021a) is not selected because the method requires manually labeled question focuses and question types.

[8] Calculation details are described in Appendix C.

Model	General Quality Metrics				General Faithfulness Metrics		Task-Specific Faithfulness Metrics				
	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	QUESTEVAL	SUMMAC	FaR	C F1	ClinicalBLEURT	MedBART	MedBERT
QFCL (Zhang et al., 2022)	0.2982	0.1121	0.2718	0.7406	0.3043	0.4072	0.4342	0.2798	0.5345	-6.8349	0.7309
PEGASUS (Zhang et al., 2020)	0.2937	0.1042	0.2743	0.7419	0.3137	0.4134	0.4291	0.2548	0.5158	-6.9167	0.7205
+ FAMESUMM (Zhang et al., 2023)	0.3045	0.1114	0.2841	0.7415	0.3086	0.4263	0.4181	0.2745	0.4923	-6.8772	0.7228
+ MEDAL (ours)	0.3155	0.1039	0.2917	0.7463	0.3238	0.4525	0.5014	0.3127	0.5154	-6.7167	0.7453
BioBART (Yuan et al., 2022)	0.3044	0.1056	0.2807	0.7494	0.3106	0.4399	0.4863	0.2813	0.5466	-7.0110	0.7359
+ FAMESUMM (Zhang et al., 2023)	0.3285	0.1244	0.3014	0.7530	0.3169	0.4722	0.5104	0.3004	0.5690	-6.9450	0.7379
+ MEDAL (ours)	0.3236	0.1341	0.3023	0.7535	0.3153	0.4958	0.5285	0.2996	0.5932	-6.8635	0.7471
Llama-2 (Touvron et al., 2023)	0.2765	0.0919	0.2516	0.7069	0.3004	0.4153	0.4446	0.2597	0.6393	-11.9865	0.6714
+ MEDAL (ours)	0.2959	0.1063	0.2987	0.7384	0.3212	0.4772	0.5173	0.2869	0.6395	-6.8926	0.7323
Med-Alpaca (Han et al., 2023)	0.3276	0.1235	0.2894	0.7490	0.3158	0.4785	0.4837	0.2721	0.5350	-6.6626	0.7356
+ MEDAL (ours)	0.3044	0.1056	0.2807	0.7494	0.3243	0.4948	0.5189	0.2665	0.5745	-6.5828	0.7367

Table 2: **Results of health question summarization.** The best results are highlighted in green, while the second-best are in blue. Baselines are reproduced using the official implementation for metric computation. We do not incorporate FAMESUMM into LLMs due to the substantial computational costs and time required.

Model	General Quality Metrics				General Faithfulness Metrics		Task-Specific Faithfulness Metrics				
	ROUGE-1	ROUGE-2	ROUGE-L	BERTSCORE	QUESTEVAL	SUMMAC	FaR	C F1	ClinicalBLEURT	MedBART	MedBERT
WGSum+CL (Hu et al., 2022b)	0.4182	0.2596	0.4007	0.7735	0.2697	0.3805	0.2826	0.2119	0.6357	-8.7590	0.7619
PEGASUS (Zhang et al., 2020)	0.4167	0.2624	0.4010	0.7788	0.2698	0.3798	0.2739	0.2128	0.6466	-8.6885	0.7584
+ FAMESUMM (Zhang et al., 2023)	0.4198	0.2896	0.4091	0.7892	0.2661	0.3756	0.2594	0.1804	0.6716	-8.8709	0.7687
+ MEDAL (ours)	0.4229	0.2797	0.3998	0.7869	0.2822	0.4176	0.3279	0.2410	0.6547	-8.1842	0.7635
BioBART (Yuan et al., 2022)	0.4091	0.2524	0.3916	0.7774	0.2782	0.4150	0.3393	0.2447	0.6341	-9.0746	0.7566
+ FAMESUMM (Zhang et al., 2023)	0.4125	0.2608	0.4085	0.7793	0.2794	0.4269	0.3289	0.2382	0.6703	-8.9820	0.7714
+ MEDAL (ours)	0.4126	0.2569	0.4088	0.7748	0.2862	0.4422	0.3710	0.2672	0.6629	-8.7749	0.7690
Llama-2 (Touvron et al., 2023)	0.2247	0.1281	0.2140	0.6423	0.2453	0.1739	0.1894	0.1473	0.7426	-21.5741	0.5620
+ MEDAL (ours)	0.2273	0.1349	0.2165	0.6459	0.2571	0.2170	0.2251	0.1922	0.7420	-19.7318	0.6379
Med-Alpaca (Han et al., 2023)	0.4111	0.2565	0.3945	0.7809	0.2638	0.3586	0.2812	0.1960	0.6407	-9.2578	0.7537
+ MEDAL (ours)	0.4160	0.2650	0.4044	0.7898	0.2767	0.3828	0.3159	0.2128	0.6503	-8.7479	0.7654

Table 3: **Results of radiology report summarization.** Baselines are reproduced using the official implementation for metric computation.

Model	ROUGE-L	FaR	C F1	ClinicalBLEURT	MedBERT
Llama-2 (Touvron et al., 2023)	0.2316	0.4434	0.2567	0.6707	0.5416
+ MEDAL (ours)	0.2408	0.4775	0.2829	0.6820	0.5498
Med-Alpaca (Han et al., 2023)	0.2314	0.2670	0.1655	0.6610	0.5195
+ MEDAL (ours)	0.2345	0.2882	0.2059	0.6559	0.5168

Table 4: **Results of doctor-patient dialogue summarization.** Baselines are reproduced using the official implementation for metric computation.

trast, MEDAL is model-agnostic and can directly optimize various summarization models.

Second, MEDAL exhibits significant improvements on SUMMAC, FaR, and C F1 metrics across all baselines, indicating that summaries generated by MEDAL include more crucial medical entities. Furthermore, the general quality metrics also show improved results after the application of MEDAL, implying an enhancement in coherence and fluency.

Improvements over large language models First, MEDAL maintains its effectiveness when dealing with summaries generated by LLMs. Notably, the performance of Llama-2 obtains significant improvement with the correction by MEDAL. Take the health question summarization as an example. There is an increase of approximately 6% on the SUMMAC metric and around 7% on the FaR metric. Furthermore, the performance of Med-Alpaca also improves further. Experiments on LLMs illustrate the capability of MEDAL to identify and rectify hallucinations in LLMs' outputs.

Second, Med-Alpaca exhibits superior performance compared to Llama-2 on health question summarization and radiology report summarization. Take radiology report summarization as an example. Med-Alpaca outperforms Llama-2 by approximately 18% on the SUMMAC metric and demonstrates an improvement of around 9% on the FaR metric. MEDAL further improves the Med-Alpaca's performance by about 2% on the SUMMAC metric and 3% on the FaR metric.

Third, for doctor-patient dialogue summarization, Med-Alpaca performs significantly worse than Llama-2, highlighting the limitations of Med-Alpaca in handling long texts. To address the context length limitation of our correction model, the source text is not directly used as input. Instead, we select relevant supporting sentences as the source text for the correction model.[9] Experimental results indicate that this approach effectively corrects erroneous content in long-text summarization, thereby enhancing the faithfulness of the summaries.

[9]The selection method is detailed in Appendix D.4.

Model	General Quality Metrics	General Faithfulness Metrics	Task-Specific Faithfulness Metrics			
	ROUGE-L	SUMMAC	FaR	C F1	ClinicalBLEURT	MedBERT
BioBART (Yuan et al., 2022)	0.2807	0.4399	0.4863	0.2813	0.5466	0.7359
+ MEDAL	0.3023	0.4958	0.5285	0.2996	0.5932	0.7471
+ MEDAL w/o SE	0.2763	0.4257	0.4763	0.2813	0.5396	0.7342
Med-Alpaca (Han et al., 2023)	0.2894	0.4785	0.4837	0.2721	0.5350	0.7356
+ MEDAL	0.2807	0.4948	0.5189	0.2665	0.5745	0.7367
+ MEDAL w/o SE	0.2802	0.4880	0.4717	0.2532	0.5435	0.7298

Table 5: **Results of the ablation study on health question summarization.** "SE" denotes self-examination.

Ratio (Positive: Negative)	ROUGE-L	SUMMAC	FaR	MedBERT
2:8	0.2916	0.4913	0.5228	0.7441
4:6 (ours)	0.3023	0.4958	0.5285	0.7471
5:5	0.3019	0.5008	0.5273	0.7359
6:4	0.2964	0.4994	0.5249	0.7493
8:2	0.3021	0.4963	0.5206	0.7357

Table 6: **Results of the robustness experiments.** We conduct experiments on health question summarization using the summaries generated by fine-tuned BioBART. "Ratio" represents the ratio of positive to negative samples in the training set.

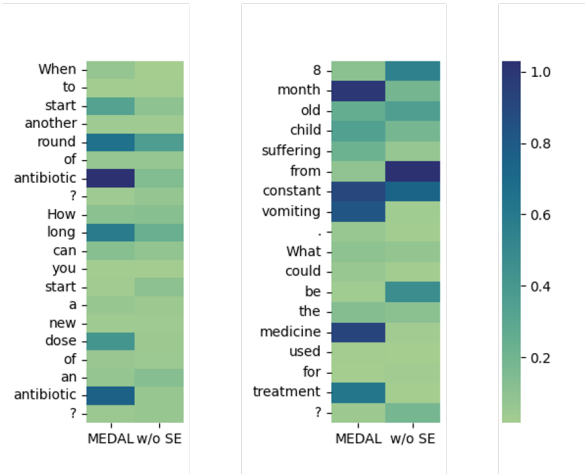

Figure 3: **Heatmap visualization of model attention.** For illustrative purposes, we focus on the attention of the first head in the final layer of the model. The heatmap depicts the attention distribution over words in the medical source text. "SE" represents self-examination.

5.2 Ablation study

Self-examination is a crucial component of our method. To substantiate its effectiveness, we conduct an ablation study.[10] The experimental results are shown in Table 5. Without self-examination, the performance of MEDAL exhibits varying degrees of decline. Remarkably, the results without self-examination are even worse than those before correction. This indicates that without self-examination, the model may incorrectly alter parts of the summary that are originally correct. Therefore, self-examination plays a pivotal role, prompting the model to consider the existence of hallucinations before rectifying the summaries.

5.3 Robustness of MEDAL

We conduct experiments to investigate the impact of the balance between positive and negative samples on the hallucination correction model. The results depicted in Table 6 indicate MEDAL's superior performance on balanced datasets (e.g., 40%, 50%, and 60% positive samples). Notably, MEDAL exhibits resilience even in scenarios of dataset imbalance. These findings highlight MEDAL's robustness to maintain efficacy across varied proportions of positive and negative samples in the training dataset.

[10]Implementation details can be found in Appendix D.5.

5.4 Further study

To delve deeper into the effectiveness of self-examination, we employ heatmap to visualize the attention matrix of the correction model. Figure 3 illustrates that the self-examination process prompts the model to allocate more attention to important medical entities. Additionally, we provide some examples of outputs from MEDAL and baselines in Appendix D.6 to show the effectiveness of MEDAL.

6 Conclusion

In this paper, we propose MEDAL, a model-agnostic post-processing framework designed to rectify hallucinations in medical summaries. MEDAL comprises a medical infilling model and a hallucination correction model. The correction model is trained using synthetic datasets generated by the infilling model. A self-examination mechanism is incorporated into the correction model to enhance its performance. Experimental results demonstrate MEDAL's superiority across

11 metrics in 3 medical text summarization tasks. MEDAL is model-agnostic and can post-process summaries to enhance their reliability. Furthermore, MEDAL exhibits potential in rectifying summaries generated by LLMs, highlighting its prospective application in the era of LLMs.

Acknowledgements

We appreciate the valuable insights provided by the anonymous reviewers. This work is supported by National Key Research and Development Program of China (No. 2022YFC3302600).

Limitations

The effectiveness of MEDAL's correction model relies on the quality and diversity of its training data, particularly the non-factual summaries tailored for hallucination correction. However, limitations arise as the errors in these non-factual summaries predominantly pertain to the entity levels and relation levels. While MEDAL has made significant improvements in mitigating hallucinations related to medical concepts, relations, and negations, it may not fully address more complex issues like contextual misunderstandings or semantic ambiguities. These complex errors pose challenges in detection due to the need for logical reasoning. Moreover, generating non-factual summaries containing these errors becomes more challenging.

Another consideration is the text length limitation of MEDAL's correction model. We select relevant supporting sentences as the source text for the doctor-patient dialogue summarization task. However, the rules of selecting relevant supporting sentences are based on ROUGE-L score and lack of the consideration of semantic information.

We will investigate new approaches capable of addressing a wider range of errors and long clinical text to improve the overall performance of medical text summarization.

Ethics Statement

Our research rigorously adheres to ethical standards by using publicly available datasets devoid of personally identifiable information. The study aims to mitigate hallucinations in medical summaries by a post-processing framework. Despite the exceptional performance of MEDAL in mitigating hallucinations, it is important to acknowledge the possibility of residual errors in its outputs. Therefore, prior to implementation, it is imperative for medical practitioners to undertake thorough assessments of faithfulness and accuracy.

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to Revise References for Faithful Summarization. In *Findings of the association for computational linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amanuel Alambo, Tanvi Banerjee, Krishnaprasad Thirunarayan, and Mia Cajita. 2022. Improving the Factual Accuracy of Abstractive Clinical Text Summarization using Multi-Objective Optimization. In *2022 44th annual international conference of the IEEE engineering in medicine & biology society (EMBC)*, pages 1615–1618.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging Linguistic Structure For Open Domain Information Extraction. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the Summarization of Consumer Health Questions. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 Shared Task on Summarization in the Medical Domain. In *Proceedings of the 20th workshop on biomedical language processing*, pages 74–85, Online. Association for Computational Linguistics.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An Investigation of Evaluation Methods in Automatic Medical Note Generation. In *Findings of the association for computational linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and others. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Songtai Dai, Quan Wang, Yajuan Lyu, and Yong Zhu. 2021. BDKG at MEDIQA 2021: System Report for the Radiology Report Summarization Task. In *Proceedings of the 20th workshop on biomedical language processing*, pages 103–111, Online. Association for Computational Linguistics.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling Language Models to Fill in the Blanks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. *arXiv preprint arXiv:2304.08247*.

Yifan He, Mosha Chen, and Songfang Huang. 2021. 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*, pages 112–118, Online. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-Rank Adaptation of Large Language Models. In *International conference on learning representations*.

Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. 2021. Word Graph Guided Summarization for Radiology Findings. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 4980–4990, Online. Association for Computational Linguistics.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022b. Graph Enhanced Contrastive Learning for Radiology Findings Summarization. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 4677–4688, Dublin, Ireland. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Sanjeev Kumar Karn, Rikhiya Ghosh, Kusuma P, and Oladimeji Farri. 2023. shs-nlp at RadSum23: Domain-Adaptive Pre-training of Instruction-tuned LLMs for Radiology Report Impression Generation. In *The 22nd workshop on biomedical natural language processing and BioNLP shared tasks*, pages 550–556, Toronto, Canada. Association for Computational Linguistics.

Sanjeev Kumar Karn, Ning Liu, Hinrich Schuetze, and Oladimeji Farri. 2022. Differentiable Multi-Agent Actor-Critic for Multi-Step Radiology Report Summarization. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1542–1553, Dublin, Ireland. Association for Computational Linguistics.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. 2021. Multi-domain Clinical Natural Language Processing with MedCAT: the Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods. *arXiv preprint arXiv:2203.05227*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W. Filice. 2019. Ontology-Aware Clinical Abstractive Summarization. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1013–1016.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa

Nakashole. 2021. 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 (volume 1: Long papers)*, pages 1505–1515, Online. Association for Computational Linguistics.

David Fraile Navarro, Mark Dras, and Shlomo Berkovsky. 2022. Few-Shot Fine-Tuning SOTA Summarization Models for Medical Dialogues. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies: Student research workshop*, pages 254–266, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP workshop and shared task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Lesley C Rink, Tolu O Oyesanya, Kathryn C Adair, Janice C Humphreys, Susan G Silva, and John Bryan Sexton. 2023. Stressors Among Healthcare Workers: A Summative Content Analysis. *Global Qualitative Nursing Research*, 10:1–13.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization Asks for Fact-based Evaluation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. 2021. Towards Clinical Encounter Summarization: Learning to Compose Discharge Summaries from Prior Notes. *arXiv preprint arXiv:2104.13498*.

Sajad Sotudeh Gharebagh, Nazli Goharian, and Ross Filice. 2020. Attend to Medical Ontologies: Content Selection for Clinical Abstractive Summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1899–1905, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, Akshay Chaudhari, and John Pauly. 2023a. RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models. In *The 22nd workshop on biomedical natural language processing and BioNLP shared tasks*, pages 449–460, Toronto, Canada. Association for Computational Linguistics.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, and others. 2023b. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. *arXiv preprint arXiv:2309.07430*.

Qianqian Xie, Jiayu Zhou, Yifan Peng, and Fei Wang. 2023. FactReranker: Fact-guided Reranker for Faithful Radiology Report Summarization. *arXiv preprint arXiv:2303.08335*.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2022. Question-Aware Transformer Models for Consumer Health Question Summarization. *Journal of Biomedical Informatics*, 128:104040.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021a. 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 (volume 2: Short papers)*, pages 249–255, Online. Association for Computational Linguistics.

Shweta Yadav, Mourad Sarrouti, and Deepak Gupta. 2021b. 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*, pages 291–301, Online. Association for Computational Linguistics.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. ACI-BENCH: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Scientific Data*, 10(1):586.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. In *Proceedings of the 21st workshop on biomedical language processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in neural information processing systems*, volume 34, pages 27263–27277.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale Medical Dialogue Datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th international conference on machine learning*, pages 11328–11339.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ming Zhang, Shuai Dou, Ziyang Wang, and Yunfang Wu. 2022. Focus-Driven Contrastive Learning for Medical Question Summarization. In *Proceedings of the 29th international conference on computational linguistics*, pages 6176–6186, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra, and Rui Zhang. 2023. FaMeSumm: Investigating and Improving Faithfulness of Medical Summarization. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 10915–10931, Singapore. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International conference on learning representations*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

A Error Types

The common error types in medical text summarization are illustrated in Table 7.

B Selection of The Lower-Ranking Range

During the inference phase, the decoding beam size is set to 15 for predicting token candidates. We select 200 medical Cloze questions and compute the ranking at which the generated tokens are inconsistent with the original tokens. Subsequently, we establish a threshold of 5 to ensure non-factual replacements occur over 95% of the time. Finally, token candidates ranked between 5 and 15 (i.e., range=[5,15]) are chosen as the non-factual replacements.

C Metrics

C.1 General quality metrics

ROUGE-1/2/L Following Zhang et al. (2023), we select ROUGE-N ($N \in \{1, 2\}$) and ROUGE-L as evaluation metrics. Lin (2004) propose ROUGE to measure the similarity between summaries. Given a candidate summary s and a reference summary r, ROUGE-N refers to the overlap of n-gram between s and r. ROUGE-L is computed based on the longest common subsequence (LCS) between s and r. We adopt the `pyROUGE`[11] to compute the ROUGE scores and select the `f_score` as the reported result.

BERTSCORE BERTSCORE computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings (Zhang* et al., 2020). We select `allenai/scibert_scivocab_cased`[12] as the BERT model to generate contextual embeddings.

C.2 General faithfulness metrics

QUESTEVAL QUESTEVAL (Scialom et al., 2021) is proposed to evaluate summarization systems without requiring reference summaries. It consists of a question generation model Q_G and a question answering model Q_A. For the question answering model Q_A, $Q_A(a|M, q)$ refers to the probability of the answer a to question q on a text M, and $Q_A(M, q)$ is the answer greedily generated from the model. For the question generation model Q_G, $Q_G(M)$ is denoted as the set of question-answer pairs (q, a) for a text M such that $Q_A(M, q) = a$. QUESTEVAL is computed as:

$$Prec(M, S) = \frac{1}{|Q_G(S)|} \sum_{(q,a) \in Q_G(S)} F1(Q_A(M, q), a) \quad (2)$$

$$Rec(M, S) = \frac{\sum_{(q,a) \in Q_G(M)} W(q, M)(1 - Q_A(\epsilon|S, q))}{\sum_{(q,a) \in Q_G(M)} W(q, M)} \quad (3)$$

$$FScore(M, S) = \frac{2 \cdot Prec(M, S) \cdot Rec(M, S)}{Prec(M, S) + Rec(M, S)} \quad (4)$$

where M and S are two sequences of tokens with M denoting the source text and S representing the corresponding evaluated summary, $W(q, M)$ is the weight of query q for text M, ϵ is the `unanswerable` token denoting that a summary contains no answer.

Following Scialom et al. (2021), the final QUESTEVAL score is the $FScore(M, S)$. We calculate it using the released package.[13]

SUMMAC SUMMAC enables NLI models to be successfully used for inconsistency detection. Laban et al. (2022) first split the source text into M blocks and split the summary into N blocks. Then, an out-of-the-box NLI model is applied to generate an $M \times N$ NLI pair matrix, where each value in the matrix is the entailment score of the source text block and the corresponding summary block. Finally, they present two models, $\text{SUMMAC}_{\text{Conv}}$ and $\text{SUMMAC}_{\text{ZS}}$. The model produces a single consistency score for a given summary by processing the pair matrix. Following Zhang et al. (2023), we adopt $\text{SUMMAC}_{\text{ZS}}$ in our evaluation by calling its package.[14]

C.3 Task-specific faithfulness metrics

FaR Shing et al. (2021) illustrate the relationship between source text, reference summary, and model-generated summary using a Venn diagram as shown in Figure 4. The Faithfulness-adjusted Recall (FaR) measures the amount of faithful and relevant information that has been included by the model-generated summary. Therefore, FaR is defined as $\frac{C}{B+C}$. Following Shing et al. (2021), we measure the rigions in Figure 4 by medical named

[11] `pyROUGE` is a Python wrapper for the ROUGE summarization evaluation package. It can be installed from https://pypi.org/project/pyrouge/

[12] https://huggingface.co/allenai/scibert_scivocab_cased

[13] https://pypi.org/project/questeval/

[14] https://pypi.org/project/summac/

Error Type	Description	Example
Entity	Generating wrong entities.	**Medical Text**: Lungs appear clear and well expanded. Previously demonstrated bilateral lower **lung airspace opacities** have resolved. **Summary**: Resolution of bilateral lower **lung opacities**.
Entity Relation	Expressing wrong relation between two entities or actions.	**Medical Text**: I thought I needed a knee repalcement. But after having a stent placed in my heart, my knee pain has been **alleviated**. What happened? **Summary**: Can a heart stent **lead** to knee pain?
Quantifier	Generating wrong dates, age, etc.	**Medical Text**: **8** month old child suffering from constant vomiting. What could be the medicine used for treatment? **Summary**: What can I give my **18** month old for constant vomiting?
Negation	Ignoring or adding negation.	**Medical Text**: Single frontal view of the chest demonstrates retrocardiac and left lung **base opacity** with a rounded contour and possible central lucency. **Summary**: **No** acute cardiopulmonary process.

Table 7: **Common error types in medical summaries.** Errors are highlighted in **red**, while evidence from the medical text used to infer the errors is highlighted in **blue**.

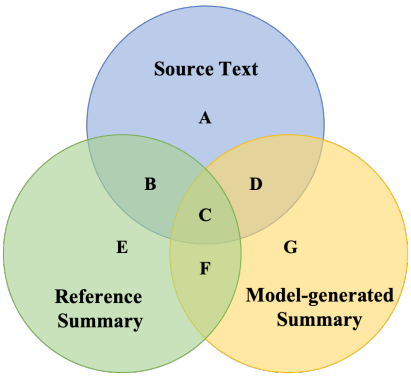

Figure 4: Relationship between source text, reference summary, and model-generated summary.

entity recognition (NER) system in `scispaCy`.[15]

Concept F1 Similar to FaR, Concept F1 is also computed based on the medical entities and we obtain the medical entities using the same method like FaR. According to the Venn diagram in Figure 4, Concept F1 is defined as:

$$\text{Concept Rec} = \frac{C+F}{B+C+E+F} \quad (5)$$

$$\text{Concept Prec} = \frac{C+F}{C+D+F+G} \quad (6)$$

$$\text{Concept F1} = \frac{2 \cdot \text{Concept Rec} \cdot \text{Concept Prec}}{\text{Concept Rec} + \text{Concept Prec}} \quad (7)$$

MedBERTScore and MedBARTScore Ben Abacha et al. (2023) update the scoring policy of two popular evaluation metrics, BERTSCORE (Zhang* et al., 2020) and BARTSCORE (Yuan et al., 2021), by providing a higher weight to the words in the summaries that have medical meaning. They identify the medical entities defined in UMLS by `MedCAT` toolkit[16]. We evaluate the summaries using the official instructions for `MedBERTScore`[17] and `MedBARTScore`[18].

ClinicalBLEURT ClinicalBLEURT relies on fine-tuning a model-based metric on a large collection of clinical notes including family medicine and orthopaedic notes. Ben Abacha et al. (2023) fine-tune BLEURT (Sellam et al., 2020) using a quality score. We use their provided fine-tuned model[19] to complete evaluation.

D Implementation Details

We implement the experiments with a single NVIDIA GeForce RTX 4090 GPU. For each task, we report the average results across 3 runs. Further implementation details can be found in our code.

D.1 Experiments on pre-trained language models

We fine-tune two pre-trained language models, namely Pegasus-large (Zhang et al., 2020) and BioBART-large (Yuan et al., 2022). The hyper-parameter settings are shown in Table 8. We select the best models based on the loss on the validation set for all tasks. Notably, these pre-trained language models are not applied to the doctor-patient dialogue summarization task because the length of

[15]The scispaCy (Neumann et al., 2019) NER identifies any span in a given text which might be an entity in UMLS (a large biomedical database) and returns the identified medical entities.

[16]https://pypi.org/project/medcat/

[17]https://github.com/abachaa/EvaluationMetrics-ACL23/tree/main/MedBERTScore

[18]https://github.com/abachaa/EvaluationMetrics-ACL23/tree/main/MedBARTScore

[19]https://github.com/abachaa/EvaluationMetrics-ACL23/tree/main/ClinicalBLEURT

Hyper-parameters	Health Question Summarization		Radiology Report Summarization	
	Pegasus-large	BioBART-large	Pegasus-large	BioBART-large
Epochs	10	10	10	10
Learning Rate	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
Train Batch Size	4	4	4	4
Eval Batch Size	4	4	4	4
Gradient Accumulation Steps	2	2	2	2
Weight Decay	0.01	0.01	0.01	0.01
Beam Size	4	4	4	4
Repetition Penalty	1.5	1.5	1.5	1.5
Length Penalty	0.8	0.8	0.8	0.8
Max Input Length	512	512	512	512
Max Output Length	84	84	256	256

Table 8: Hyper-parameter settings of pre-trained language models.

Use	Prompt Component
Health Question Summarization	Summarize the patient health question into one question of 15 words or less.
Radiology Report Summarization	Summarize the radiology report findings into an impression with minimal text.
Doctor-Patient Dialogue Summarization	Summarize the patient/doctor dialogue into an assessment and plan.
Prefix	You are a knowledgeable medical professional.

Table 9: **Model prompts.** The prefix is applied to every task.

the doctor-patient dialogues exceeds the context length of Pegasus-large and BioBART-large.

D.2 Experiments on LLMs

Following Van Veen et al. (2023b), the prompts for medical summarization tasks are depicted in Table 9. We fine-tune LLMs with LoRA (Hu et al., 2022a) using the Adam optimizer. Table 10 displays the hyper-parameters for fine-tuning in different tasks. The final model used for testing is obtained at the end of training, based on the specified number of epochs.

D.3 Experiments on MEDAL

For the medical infilling model $\mathcal{M}_I$ and the hallucination correction model $\mathcal{M}_C$, we select the best models based on the loss on the validation set. The hyper-parameter settings are shown in Table 11 and 12.

D.4 Selection of relevant supporting sentences

To adapt MEDAL to doctor-patient dialogue summarization, we select the most relevant sentences from the source text as input context instead of using the entire source text. For each sentence i in the model-generated summary, we choose a sentence from the source text that yields the highest ROUGE-L score computed with i. These selected sentences are then concatenated as the input source text.

D.5 Ablation study

We show the input and output formats of the correction model in Table 13.

Besides, we implement further experiments to explore whether clear instructions can help the model perform better. The experiment results are shown in Table 14. The clear instructions indeed improve model performance.

D.6 Examples of outputs

Figure 5 illustrates MEDAL's effectiveness in correcting errors found in model-generated summaries. In the first two examples, the original outputs generated by summarization model contains the wrong entities and quantifiers. MEDAL accurately captures the key medical entities and generates a complete summary. Similarly, in the last example, Pegasus overlooks the negation "no", whereas MEDAL generates the correct summary. These examples highlight MEDAL's capability to correct hallucinations within summaries. We design medical Cloze questions based on common errors and employ an infilling model to generate non-factual summaries for training the correction model. Additionally, we incorporate a self-examination mechanism into the correction model, further enhancing its performance.

Hyper-parameters	Health Question Summarization		Radiology Report Summarization		Doctor-Patient Dialogue Summarization	
	Llama-2	Med-Alpaca	Llama-2	Med-Alpaca	Llama-2	Med-Alpaca
Epochs	10	10	3	3	10	10
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Gradient Accumulation Steps	4	4	4	4	4	4
Train Batch Size	8	8	8	8	4	4
Eval Batch Size	4	4	4	4	2	2
Beam Size	1	1	1	1	1	1

Table 10: Hyper-parameter settings of large language models.

Hyper-parameters	Health Question Summarization	Radiology Report Summarization	Doctor-Patient Dialogue Summarization
Epochs	10	10	10
Learning Rate	1×10^{-5}	1×10^{-5}	1×10^{-5}
Batch Size	8	8	8
Train Beam Size	1	1	1
Eval Beam Size	15	15	15
Max Input Length	512	512	512

Table 11: Hyper-parameter settings of the medical infilling model $\mathcal{M}_I$.

Hyper-parameters	Health Question Summarization	Radiology Report Summarization	Doctor-Patient Dialogue Summarization
Epochs	10	10	10
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Train Batch Size	4	4	4
Eval Batch Size	4	4	4
Weight Decay	0.01	0.01	0.01
Beam Size	4	4	4
Repetition Penalty	1.5	1.5	1.5
Length Penalty	0.8	0.8	0.8
Max Input Length	512	512	512
Max Output Length	128	128	300

Table 12: Hyper-parameter settings of the hallucination correction model $\mathcal{M}_C$.

	Phase	Input Sequence	Target Sequence
Flan-T5	Training Phase	Positive Sample: [SOURCE] <SEP> [REFERENCE] Negative Sample: [SOURCE] <SEP> [NON-FACTUAL]	[REFERENCE] [REFERENCE]
	Inference Phase	[SOURCE] <SEP> [MODEL-GENERATED]	[CORRECTED]
+ self-examination	Training Phase	Positive Sample: [PROMPT]. Text: [SOURCE]. Summary: [REFERENCE]. Negative Sample: [PROMPT]. Text: [SOURCE]. Summary: [NON-FACTUAL].	[NO] [YES]. [REFERENCE]
	Inference Phase	[PROMPT]. Text: [SOURCE]. Summary: [MODEL-GENERATED].	[NO] / [YES]. [CORRECTED]

Table 13: **Input and output formats of models.** `[SOURCE]` represents the source text m. `[REFERENCE]` corresponds to the reference summary r. `[NON-FACTUAL]` denotes our generated summary produced by the infilling model $\mathcal{M}_I$. `[MODEL-GENERATED]` is the model-generated summary s'. `[CORRECTED]` represents the corrected summary s generated by the correction model $\mathcal{M}_C$. `[YES]` and `[NO]` are the `self-examination` tokens introduced by us. `[PROMPT]` signifies the prompt instruction "Check for any hallucination in the summary, and if found, correct them."

Model	ROUGE-L	SUMMAC	FaR	C F1	ClinicalBLEURT	MedBERT
BioBART(Yuan et al., 2022)	0.2807	0.4399	0.4863	0.2813	0.5466	0.7359
+MEDAL w/ SentInst	0.3146	0.4967	0.5259	0.3049	0.5943	0.7490
+MEDAL (ours)	0.3023	0.4958	0.5285	0.2996	0.5932	0.7471
Med-Alpaca(Han et al., 2023)	0.2894	0.4785	0.4837	0.2721	0.5350	0.7356
+MEDAL w/ SentInst	0.2961	0.5084	0.5177	0.2737	0.5843	0.7362
+MEDAL (ours)	0.2807	0.4948	0.5189	0.2665	0.5745	0.7367

Table 14: **Results of the ablation study on instructions.** w/ SentInst denotes that we replace `[YES]` with "This summary contains hallucination, and here is the revision:" and `[NO]` with "This summary contains no hallucination.".

Entity-Level Errors

Source	Outbreak of **red raise sores on back shoulder and arms** from shoulder to wrist with raised bulbs of liquid. treatment and maintenance?
MedAlpaca	What are the treatments for **herpes simplex virus skin lesions**?
+ MEDAL	What are the treatments for **red sores on back shoulder and arm** with raised bulbs of liquid?

Quantifier Errors

Source	Recently my son took an eye examine and he had 20/40 in one eye and **20/25** in the other would he need glasses? Then when they continue the eye examine they said the Fundus could not be ……
Pegasus	What are the treatments for a child with 20/40 in one eye and **25/25** in the other?
+ MEDAL	What are the treatments for a child with 20/40 in one eye and **20/25** in the other?

Negation Errors

Source	The heart is normal size with normal appearance of the cardia mediastinal silhouette. There are **no degenerative changes and thoracic spine**.
Pegasus	Normal heart size. **Thoracic spine degeneration**.
+ MEDAL	Normal heart size. **No thoracic spine degeneration.**

Figure 5: **Three examples to show the effectiveness of MEDAL.** We mark the medical terms in **blue** and the errors in **red**.