Aligning Factual Consistency for Clinical Studies Summarization through Reinforcement Learning

Xiangru Tang[♠] **Arman Cohan**[♠] **Mark Gerstein**[♠]
[♠] Yale University, New Haven, CT 06520, USA

{xiangru.tang, arman.cohan, mark.gerstein}@yale.edu

Abstract

In the rapidly evolving landscape of medical research, accurate and concise summarization of clinical studies is crucial to support evidence-based practice. This paper presents a novel approach to clinical studies summarization, leveraging reinforcement learning to enhance factual consistency and align with human annotator preferences. Our work focuses on two tasks: Conclusion Generation and Review Generation. We train a CONFIT summarization model that outperforms GPT-3 and previous state-of-the-art models on the same datasets and collects expert and crowd-worker annotations to evaluate the quality and factual consistency of the generated summaries. These annotations enable us to measure the correlation of various automatic metrics, including modern factual evaluation metrics like QAFactEval, with human-assessed factual consistency. By employing top-correlated metrics as objectives for a reinforcement learning model, we demonstrate improved factuality in generated summaries that are preferred by human annotators.

1 Introduction

Recently, the exponential growth of medical literature, specifically in the realm of clinical studies such as randomized controlled trials (RCTs), has underscored the necessity for efficient summarization techniques (Cohan et al., 2018; Sotudeh Gharebagh et al., 2020; Guo et al., 2021; Wang et al., 2020; Luo et al., 2022). Clinicians and researchers face the arduous task of sifting through vast amounts of information daily to remain abreast of the latest findings and advancements in their respective fields (Abacha et al., 2021; Chaves et al., 2022). Summarizing clinical studies enables healthcare professionals to access crucial information more rapidly, ensuring that their decisions and treatment plans are informed by the most recent, evidence-based knowledge. As a result, the development of effective and accurate summarization techniques for clinical studies has become an essential area of research in the medical domain (Shieh et al., 2019a; Wang et al., 2021; Wallace et al., 2021; DeYoung et al., 2021; Xie et al., 2022; Otmakhova et al., 2022a; Tang et al., 2023).

Automatic summarization of clinical studies is fundamental for systems that aim to interpret the vast array of available medical literature (Shieh et al., 2019a; Sotudeh Gharebagh et al., 2020; Otmakhova et al., 2022b; Tangsali et al., 2022). Randomized controlled trials (RCTs) are considered the gold standard of clinical evidence among various study types, including cohort studies, observational studies, and case studies (Concato et al., 2000; Katsimpras and Paliouras, 2022). The ability to efficiently process and summarize the massive volume of RCTs holds great potential for enhancing clinical decision-making (Meldrum, 2000; Ramprasad et al., 2023).

To delve deeper into clinical study summarization, we simultaneously explore single-document and multi-document summarization techniques. For single-document summarization, we propose an RCT conclusion generation task based on the PubMed 200k RCT sentence classification dataset (Dernoncourt and Lee, 2017). We utilize the PubMed RCT200k dataset (Dernoncourt and Lee, 2017), with original annotations for concluding sentences, meaning our summarization system's objective is to generate concluding sentences for a clinical study. In the case of multi-document summarization, we examine the challenge of automatically generating a narrative biomedical summary from multiple trial reports. Here inputs are titles and abstracts from systematic reviews previously conducted by members of the Cochrane collaboration[1] (Wallace et al., 2021), using the review abstract as our target, shown as Figure. 1.

Ensuring the factual consistency of summaries is vital in the medical field, as they must precisely

[1] https://www.cochrane.org/

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 48–58

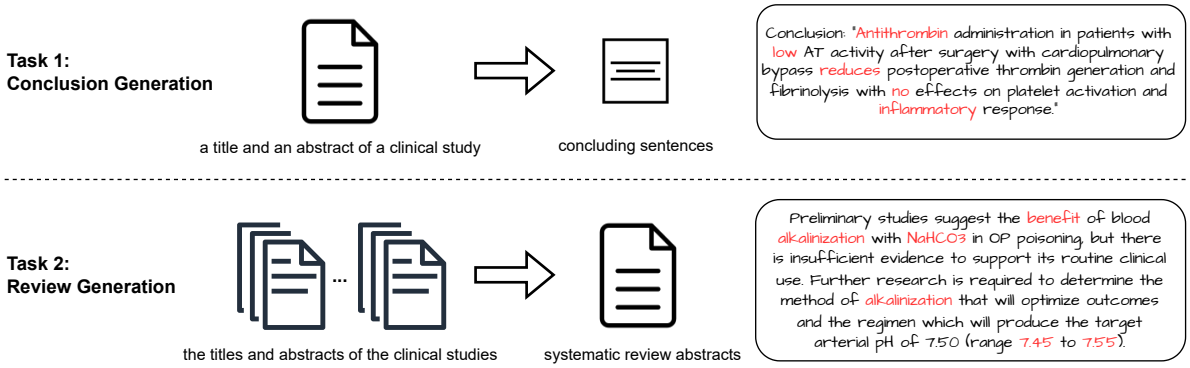

Figure 1: The diagram provides an overview of our tasks, which include single-document summarization and multi-document summarization. For the single-document summarization task, the input consists of a title and an abstract of a clinical study, with the goal being to generate concluding sentences. On the other hand, the input for the multi-document summarization task consists of the titles and abstracts from a corresponding review. Highlighted in red are various specialized medical concepts, logical reasoning, and numerical understanding, which introduce new challenges for clinical study summarization.

convey evidence to readers who make decisions for real patients. Wallace et al. demonstrated that modern summarization systems often struggle to create factually consistent summaries and tend to generate content with factual discrepancies compared to the input. At the same time, traditional automatic evaluation metrics have been deemed insufficient for assessing correctness, leading to a reliance on human evaluation for verifying generated summaries (Kryscinski et al., 2020; Maynez et al., 2020; Xie and Wang, 2023). However, such human evaluation demands medical expertise, which can be both expensive and challenging to scale. To tackle this issue, our work focuses on evaluating various automated metrics for their correlation with factual consistency and improving the factual consistency of clinical study summarization systems. In our approach, we utilize the top-correlated metrics from the previous experiment as the objective for a reinforcement learning (RL) model, like previous work (Paulus et al.). By doing so, we aim to guide the model toward generating more factually consistent summaries. Our results show that the RL-based models exhibit improved factuality and are preferred by human annotators, demonstrating the effectiveness of using RL for enhancing factual consistency in clinical study summarization systems.

Our main contributions: We emphasize our focus on clinical studies and discuss the unique challenges associated with their summarization. Our experiments feature comprehensive benchmarks and modern factual evaluation metrics, such as QAFactEval (Fabbri et al., 2022). We gathered annotations from both crowd workers and domain experts to assess the factual correctness of summaries generated by state-of-the-art models. By utilizing the top-correlated metrics as the objective for a reinforcement learning (RL) model, our results demonstrate improved factuality that is preferred by human annotators, showcasing the effectiveness of our approach.

2 Related Work

2.1 Clinical Trial Summarization

Clinical trial summarization has emerged as an important area of research due to the increasing volume of medical literature and the need for efficient information extraction. Early clinical trial summarization techniques often employed rule-based and template-based approaches, which relied on predefined templates and hand-crafted rules to generate summaries. For example, Demner-Fushman and Lin utilized a rule-based system to extract PICO elements from clinical trial abstracts. However, these methods were limited by their reliance on predefined templates and rules, which made them less adaptable to various domains and less effective in capturing the nuances of clinical trials. As machine learning gained traction, researchers began to explore feature-based approaches for clinical trial summarization. For instance, (Shieh et al., 2019b) worked towards understanding medical randomized controlled trials by conclusion genera-

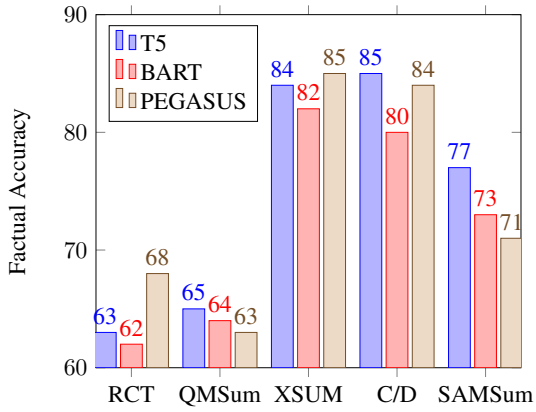

Figure 2: Bar chart illustrating the factual accuracy scores (higher is better) for different text summarization models (T5, BART, PEGASUS) on various datasets. The datasets are represented by the following abbreviations: RCT (RCT200K), QMSum (QMSum), XSUM (XSUM), C/D (CNNDM), and SAMSum (SAMSum).

tion. (Wallace et al., 2021) generated narrative summaries of RCTs with neural multi-document summarization. Although these methods showed promise, they still required significant manual feature engineering and were sensitive to the choice of features. The advent of deep learning has led to substantial improvements in clinical trial summarization. Neural network-based models, such as sequence-to-sequence models, have been employed for summarizing clinical trials. For example, (DeYoung et al., 2020) presented Evidence Inference 2.0, which focused on more data and better models in the biomedical domain. Additionally, (DeYoung et al., 2021) introduced MS2, a multi-document summarization approach for medical studies. These studies demonstrated superior performance compared to traditional machine learning methods.

2.2 Factual Consistency in Summarization

Factual consistency is a critical aspect of text summarization, as it ensures that generated summaries accurately represent the source content (Maynez et al., 2020). Previous works have discussed the challenges associated with achieving factual consistency, including issues like hallucination (Zhang et al., 2022a; Sridhar and Visser, 2022; van der Poel et al., 2022), and various techniques employed to address these challenges, such as reinforcement learning (Wan and Bansal, 2022a) and model fine-tuning (Zhang et al., 2022a; Wan and Bansal, 2022b; Tang et al., 2022b). Numerous existing models have attempted to address this issue, including extractive (Zhang et al., 2022b), abstractive (Ladhak et al., 2022; Chen et al., 2021; Wan and Bansal, 2022b). Additionally, several researchers have proposed better evaluation metrics to assess factual inconsistency, such as QAFactEval (Fabbri et al., 2022) and FactCC (Kryscinski et al., 2020). Promising avenues for future research may utilize high-quality negative examples (Wang et al., 2022), better evaluation metrics (Luo et al., 2023), and novel model architectures (Chaudhury et al., 2022) to improve the factual consistency of generated summaries. And recently, there has been a growing interest in the generation of surveys (Li et al., 2021). However, there is currently limited discussion concerning factual consistency issues in clinical survey or studies summarization.

3 Preliminarily

3.1 Datasets and Formulation of Tasks

We present two tasks:

Task 1: Conclusion Generation. We employ a modified version of the PubMed 200k RCT dataset, initially designed for sequential sentence classification. This dataset emphasizes medical abstracts, particularly randomized controlled trials (RCTs), which are considered the most reliable source of medical evidence. Each sentence in the dataset is labeled with a specific class that corresponds to the section it originates from Objective, Background, Conclusions, Methods, and Results. For inclusion, abstracts must (1) pertain to an RCT and (2) be structured. Among the 195,654 abstracts meeting both criteria, we allocate them into training (190,654), validation (2,500), and testing (2,500) sets. We assemble sentences from the Objective, Background, Methods, and Results classes of each abstract into a single paragraph, which serves as the input text for summarization. Sentences marked as Conclusions function as the reference summary for the medical abstract.

Task 2: Review Generation. Our dataset comprises systematic review abstracts as well as the titles and abstracts of the clinical trials summarized within these reviews. All data is sourced from PubMed, exclusively using abstracts. On average, each review encompasses ten trials, featuring an average abstract length of 245 words. We employ the "authors' conclusions" subsection of the systematic review abstract as our target summary (with an average of 75 words). The dataset is randomly partitioned into training (3,619 reviews), development

(455 reviews), and testing (454 reviews) sets.

Additionally, focusing on clinical trials in comparison to other summarization tasks, such as news, meetings, dialogues, movies, emails, sports, or games, is driven by the unique challenges presented in scientific and clinical summarization. Clinical trial summaries often require more logical reasoning, numerical understanding, and the accurate representation of a vast array of domain-specific terminology. The importance of correctly conveying specific drug values, concentrations, and scales adds significant complexity to summarization models. In an evaluation we conducted, we crowdsourced 100 different summarization examples (including CNNDM (Hermann et al., 2015), XSUM (Narayan et al., 2018), SAMSum (Gliwa et al., 2019), QMSum (Zhong et al., 2021), and RCT200K) and applied state-of-the-art summarization models, such as Pegasus (Zhang et al., 2020), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020). We asked the participants to assess the factual consistency of the generated summaries and found that clinical summarization exhibited the lowest factuality accuracy (highest error rate). This highlights the challenge and importance of focusing on clinical trial summarization, ensuring accuracy and consistency in conveying critical information.

3.2 Meta evaluation

First, we assess the performance of automatic metrics using 200 test summaries for two tasks, in Table. 1 and Tabel. 2. We examine the results obtained using fine-tuned PEGASUS (Zhang et al., 2020), CONFIT (Tang et al., 2022b), BART models, as well as zero-shot GPT-3 summaries. Here, we use GPT-3 davinci model. ConFiT introduces a novel training approach that enhances the factual accuracy and overall quality of summaries through contrastive fine-tuning, emphasizing error identification. Although the original study focused on dialogue summarization, we adapt and fine-tune the approach for clinical summarization.

Additionally, we engage both domain experts and general crowd-workers to evaluate the generated summaries. Following the methodology of (Tang et al., 2022a), we employ a 10-point Likert scale for expert and crowd-worker annotators to assess factual consistency. For each summary, we have two crowd-workers and one expert providing scores, and we take the average of these three ratings. We also adopt a block design in our evaluation process: each crowd-worker evaluates 10 summaries, while we engage two MD students with medical bachelor's degrees, who each assess 100 samples. We employed the AWS Mechanical Turk (MTurk) platform to engage crowd-workers for our task. Each crowd-worker annotator received a $10 compensation. We recruited 40 MTurk workers with strong track records, using these qualifications: a HIT approval rate of at least 99%, a minimum of 200 approved HITs, and residence in one of the following English native-speaking countries: US, Australia, Canada, New Zealand, or UK.

Table. 1 displays the performance of the summarization systems evaluated using automatic reference-based evaluation metrics, such as ROUGE, METEOR, and BLEU, highlighting the differences in scores between the models for each task. The model with the highest scores across most metrics for Task 1 is `CONFIT`, while for Task 2, it is again `CONFIT`. Notably, `CONFIT`outperforms the other models in terms of ROUGE, METEOR, and EM and F1 of QAEval.

Table. 2 showcases the performance of the same summarization systems evaluated using automatic reference-free metrics, including SUPERT(Gao et al., 2020), BLANC (Vasilyev et al., 2020), QuestEval (Scialom et al., 2021), QAFactEval (Fabbri et al., 2022), FactCC (Kryscinski et al., 2020), DAE (Goyal and Durrett, 2021), and SummaC (Laban et al., 2022). This table provides an alternative perspective on the performance of these models, as it does not rely on reference summaries for evaluation. However, PEGASUS has the highest SUPERT score and the highest QuestEval score. Meanwhile, for Task 2, PEGASUS again scores the highest in SUPERT, while BART achieves the highest QuestEval score.

Table. 3 presents the instance-level Pearson correlation of various metrics with human factual consistency ratings on Task 1 and Task 2. This table helps identify which metrics are more strongly correlated with factual consistency, providing insights into the most reliable evaluation methods for measuring the factuality of generated summaries. We can observe that DAE (0.5743 for Task 1 and 0.2089 for Task 2) and QAFactEval (0.5516 for Task 1 and 0.3147 for Task 2) have the highest correlation with human consistency ratings. This indicates that DAE and QAEval are more closely aligned with human judgment when evaluating factual consistency for clinical studies summarization.

Dataset	Model	Overlap-Based			QAEval	
		ROUGE(1/2/L)	METEOR	BLEU	EM	F1
Task 1	PEGASUS	34.96/14.75/28.35	.23	7.2	.104	.159
	CONFIT	38.55/17.15/31.52	.32	6.5	.136	.210
	BART	35.10/13.90/28.52	.24	5.8	.098	.162
	GPT-3	31.92/11.38/24.77	.24	3.7	.097	.158
Task 2	PEGASUS	33.15/13.60/26.90	.21	.13	.094	.155
	CONFIT	37.40/16.50/30.40	.30	.15	.136	.202
	BART	25.67/9.55/21.47	.19	.03	.072	.126
	GPT-3	27.48/10.72/22.24	.21	.04	.084	.143

Table 1: Performance of various summarization systems evaluated using automatic reference-based evaluation metrics.

Dataset	Model	Overall Quality		Factuality (QA-based)		Factuality (NLI-based)		
		SUPERT	BLANC	QuestEval	QAFactEval	FactCC	DAE	SummaC
Task 1	PEGASUS	.5475	.0615	.7380	4.4095	.3750	.8235	.1145
	CONFIT	.5596	.0813	.7343	3.8354	.1823	.7587	-.0521
	BART	.5348	.0564	.7803	3.7537	.2021	.7568	-.0594
	GPT-3	.5579	.0751	.7268	3.6419	.2438	.6682	-.0719
Task 2	PEGASUS	.6292	.1144	.7129	4.2141	.7223	.7966	.2425
	CONFIT	.5910	.0910	.7360	3.6820	.2510	.7410	.0110
	BART	.5443	.0658	.7529	3.5814	.2837	.7382	.0279
	GPT-3	.5411	.0606	.7173	3.2352	.3998	.6574	-.0711

Table 2: Performance of various summarization systems evaluated using automatic reference-free metrics.

Metric	Task 1	Task 2
ROUGE(1/2/L)	0.2721	0.0812
METEOR	0.2243	0.1309
BLEU	0.2567	0.1548
QAFactEval	0.5516	0.3147
SUPERT	0.2945	0.1157
BLANC	0.3304	0.0846
QAEval	0.2102	0.1261
QuestEval	0.5337	0.3816
FactCC	0.3719	0.1675
DAE	0.5743	0.2089
SummaC	0.3621	0.2953

Table 3: Instance-level Pearson correlation of various metrics with factual consistency ratings on Task 1 and Task 2.

Furthermore, we observe that automatic metrics report notably lower results for GPT-3 summaries compared to fine-tuned models in both of our tasks. However, in our manual evaluation, the performance of GPT-3 is actually very high, surpassing the other three models. This indicates that GPT-3 excels in factual consistency. Nonetheless, the results of automatic indicators, whether they measure overall quality or factuality evaluation, are entirely opposite to those of manual evaluation. This leads us to believe that automatic metrics may not be reliable for comparing the quality of zero-shot summaries. The evaluation method for zero-shot summaries should probably differ from that of fine-tuned summaries, as it may be more subjective. We plan to investigate this issue further in future research.

4 Methodology

We have observed that DAE, QuestEval, and QAFactEval exhibit high correlations with factual consistency across the two datasets. Therefore, our goal is to incorporate DAE and QAFactEval as reinforcement learning objectives to enhance the performance of the base model in text summarization. To achieve this, we can augment the base model's loss function with QAFactEval using reinforcement learning, specifically the policy gradient approach.

Consider the base model's loss function $L(\theta)$, where θ represents the model's parameters. The reinforcement learning objective, e.g. using QAFactEval, is to maximize the reward function $J(\theta)$ for each trajectory τ under the policy θ. The reward function can be defined as:

$$J(\theta) = \mathbb{E}_{\tau \sim p\theta(\tau)}[R(\tau)] \quad (1)$$

Model	Human	QAFactEval	DAE
CONFIT	6.3	3.8350	0.7585
+RL QuesEval	7.1	3.8365	0.7602
+RL QAFactEval	7.3	3.8493	0.7634
+RL DAE	7.3	3.8375	0.7748

Table 4: Evaluation results on Task 1 Conclusion Generation.

Model	Human	QAFactEval	DAE
CONFIT	6.3	3.6820	0.7410
+RL QuesEval	7.0	3.6855	0.7435
+RL QAFactEval	7.3	3.6929	0.7442
+RL DAE	6.3	3.6837	0.7514

Table 5: Evaluation results on Task 2 Review Generation.

Here, $R(\tau)$ denotes the QAFactEval reward for a given trajectory τ, and $p_\theta(\tau)$ represents the probability of the trajectory under the policy θ. To incorporate the QAFactEval reward into the base model's loss function, we can modify the original loss function $L(\theta)$ as follows:

$$L'(\theta) = L(\theta) - \alpha J(\theta) \tag{2}$$

In this new loss function $L'(\theta)$, α is a hyperparameter that balances the contribution of the original loss function and the reinforcement learning objective. By optimizing this new loss function, the base model can generate summaries that better align with the QAFactEval metric.

To approximate the expected reward $J_t(\theta)$ at each time step using a single sample, we can employ the following Monte Carlo estimation:

$$\mathbb{E}_{\tau \sim p\theta(\tau)}[R_t(\tau)] \approx \frac{1}{M} \sum_{i=1}^{M} r_{i,t} \tag{3}$$

Here, $r_{i,t}$ denotes the QAFactEval reward for a single trajectory τ_i at time step t, and M is the number of samples (trajectories) used to approximate the expected reward. Using this approximation, we can update the BART model's loss function at each time step as follows:

$$L'_t(\theta) = L_t(\theta) - \alpha \frac{1}{M} \sum_{i=1}^{M} r_{i,t} \tag{4}$$

In this new loss function $L'_t(\theta)$, α is a hyperparameter that balances the contribution of the original loss function and the reinforcement learning objective for each time step. By optimizing this new loss function at each step, the base model can generate summaries that better align with the QAFactEval metric at every time step.

5 Experiment and Results

5.1 Setting

For the models of PEGASUS, CONFIT, and BART, learning rate was set to 3e-5, a dropout rate of 0.1 was used, a batch size of 32, and GPT-3 we use OpenAI API. And here we have beam search for decoding with beam size 3. We use their original code base for all metrics [2].

5.2 Results

Here we present our experimental results, which demonstrate the effectiveness of incorporating reinforcement learning (RL) objectives into our base models, ConFiT. We use three different metrics, QuesEval, QAFactEval, and DAE, as RL objectives to improve the models' performance in generating factually consistent summaries. Tables 4 and 5 show the results for Task 1 and Task 2, respectively. As the tables show, for both tasks, the models augmented with RL objectives exhibit improved performance across all three metrics. This indicates that incorporating reinforcement learning objectives using QuesEval, QAFactEval, and DAE successfully improves the factual consistency of the generated summaries. It's worth noting that the performance improvement was consistently observed and it did not merely result from some testing of statistical significance. We conducted multiple experiments to ensure the stability and reliability of these performance gains.

In this study, we adopt the same human evaluation setup as previously mentioned, using a 0-10 point scale for rating the generated summaries. This consistent evaluation approach allows us to effectively compare the performance of our models and assess their ability to generate factually consistent summaries in both tasks.

In Table 6, we provide the original text, reference summary, and summaries generated by three different models. It's important to clarify that this example was not cherry-picked. It is representative and indicative of the general trends observed in our data, rather than being an exceptional case chosen to support our argument. From the original

[2]https://github.com/salesforce/QAFactEval, https://github.com/salesforce/factCC, and https://github.com/tingofurro/summac.

	Study Details and Results
Study 1	Antiemetic activity of ondansetron in acute gastroenteritis.,"The mechanism of nausea and vomiting associated with gastroenteritis is unknown. The role of 5-HT3 receptors in emesis associated with gastroenteritis was investigated in paediatric patients. A randomized, double-blind, placebo-controlled, parallel-group study was conducted in three groups of 12 patients each, receiving either a single i.v. dose of ondansetron (0.3 mg/kg), metoclopramide (0.3 mg/kg) or placebo (sterile saline). Food was restricted and oral rehydration was administered for 4 h. During 0-24 h, the number of emetic episodes experienced was significantly greater with placebo (mean = 5) than ondansetron (mean = 2), (P = 0.048), and the proportion of patients experiencing no emesis was significantly greater with ondansetron (58%) than placebo (17%). A numerical difference, in favour of ondansetron, was observed between ondansetron and metoclopramide groups for both of the above parameters. Fewer treatment failures were observed with ondansetron (17%) than placebo (33%) and metoclopramide (42 %). More diarrheal episodes were observed in the groups receiving anti-emetic treatment. Therefore, serotonin, acting through 5HT3 receptors, may play a role in this form of emesis." 364,CD005506,19758398,Clinical trial: oral ondansetron for reducing vomiting secondary to acute gastroenteritis in children–a double-blind randomized study...
Study 2	Oral ondansetron for gastroenteritis in a pediatric emergency department...
Study 3	All subjects also received intravenous NS at 10-20 mL/kg/hr...
Study 4	Dimenhydrinate in children with infectious gastroenteritis...
Study 5	Hospital admission occurred in nine patients (20.5%) receiving placebo...
Study 6	...Ondansetron, a 5-HT3 receptor antagonist marketed as Zofran...
Study 7	...with a clinically important difference of at least 30% when comparing 2 groups of patients presenting...
Reference	Oral ondansetron increased the proportion of patients who had ceased vomiting and reduced the number needing intravenous rehydration and immediate hospital admission. Intravenous ondansetron and metoclopramide reduced the number of episodes of vomiting and hospital admission, and dimenhydrinate as a suppository reduced the duration of vomiting.
CONFIT	A randomized, double-blind, placebo-controlled study investigated the role of 5-HT3 receptors in emesis associated with gastroenteritis in pediatric patients. Results showed that oral ondansetron significantly reduced the number of emetic episodes experienced compared to placebo and increased the proportion of patients experiencing no emesis (58% vs 17%, P = 0.048).
CONFIT with RL	These articles are randomized controlled trials that investigate the efficacy and safety of ondansetron compared to placebo or other antiemetic medications in reducing vomiting episodes and improving oral rehydration therapy. The studies show that ondansetron is effective in reducing vomiting episodes and improving oral rehydration in children with gastroenteritis, and it is well-tolerated with few adverse effects.
GPT3	The studies presented in this table provide evidence that ondansetron, a 5-HT3 receptor antagonist, is effective in reducing vomiting episodes and improving oral rehydration in children with gastroenteritis. The use of ondansetron leads to fewer treatment failures, hospital admissions, and the need for intravenous rehydration. Overall, these results suggest that ondansetron and dimenhydrinate can be useful adjuncts to oral rehydration therapy in managing gastroenteritis in pediatric patients.

Table 6: We provide the original text, reference summary, and summaries generated by three different models, including ConFIT, ConFIT with reinforcement learning (RL) incorporated into its loss function, and GPT-3. The original text comprises various clinical studies; however, due to its length, we have omitted some portions of the input documents.

text marked in blue, we can see that the p-value of 0.048 refers to the comparison between the placebo and ondansetron groups in terms of the number of emetic episodes experienced, rather than the proportion of patients experiencing no emesis. However, in the original output from the ConFIT model, the proportion is incorrectly mixed with the wrong p-value, which we have marked in red. In contrast, the summaries generated by the models with reinforcement learning incorporated into their loss functions do not exhibit any factual errors.

6 Future Work

We find that GPT-3-generated summaries, though scoring poorly on automatic evaluation metrics, were considered superior in quality through manual evaluation. Future studies could focus on this to offer clearer insights and establish a stronger basis for this assertion.

This paper also motivates a potential avenue for improvement, suggesting a shift from traditional automatic metrics towards utilizing advanced models like GPT-4 as evaluators. The premise behind this is to test the correlation between the scores assigned by GPT-4 and human evaluators. If a high correlation exists (which we hypothesize might be the case), it would be feasible to employ GPT-4 scores to guide the tuning process in reinforcement learning. This approach essentially distills the capabilities of GPT-4, leveraging its advanced understanding and evaluative capacity to enhance the performance and efficiency of the summarization models. This proposition could herald a novel direction in the evaluation and tuning of such models, potentially offering more reliable and nuanced performance metrics.

7 Conclusion

In the field of clinical studies summarization, there has been limited research on factual consistency so far. We have demonstrated that, for single-document and multi-document summarization of clinical studies, there are two main issues: 1) the accuracy of automatic metrics for evaluating factual consistency is limited, and 2) existing models have their limitations. We employed human evaluation for assessing factual consistency, and this analysis has been conducted over a larger set of automatic metrics to provide a more comprehensive picture. Furthermore, we demonstrate that further optimizing the model using reinforcement learning (RL) with the metric as a reward can result in significant improvements in factual consistency. Our contributions include a simple yet effective approach for two medical summarization tasks, validation of several automatic evaluation metrics for their correlation with expert-assessed factualness, and the identification of the best-correlating metric to guide generation models toward enhanced summary correctness. This work lays the foundation for the development of more robust clinical trial summarization systems, facilitating the efficient dissemination of medical knowledge to practitioners and researchers.

8 Limitation

While this study provides valuable insights into the performance of summarization models across various domains, there are several limitations that should be noted. Primarily, it is clear from the results that these models exhibit poor performance on RCT compared to other domain datasets. However, it should be noted that this performance gap is likely due, at least in part, to the fact that these models were not trained on medical documents. The complexity of medical terminology and its syntax often requires specific knowledge and understanding that general language models might not possess. Thus, it might not be entirely fair to infer that these summarizers find clinical summarization inherently challenging based on this data alone. In order to address this limitation, it is recommended that future research should involve the same experiments using model checkpoints that have been finetuned on medical text data.

Another notable limitation of the study revolves around the incremental improvements shown by the summarizers for the evaluation metrics used in the reward function. While it is encouraging to observe these slight improvements, it's important to question and validate whether these changes truly signify an enhancement in the model's factuality. It's plausible that the training focused primarily on improving factuality might inadvertently compromise other aspects of the generated text, such as its fluency or ROUGE scores. To gain a more comprehensive understanding, it would be valuable to conduct additional experiments and analyses. This comprehensive evaluation is critical to gain a more nuanced understanding of the trade-offs involved in model training and optimization.

References

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.

Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, and Alexander Gray. 2022. X-FACTOR: A cross-metric evaluation of factual correctness in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7100–7110, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrea Chaves, Cyrille Kesiku, and Begonya Garcia-Zapirain. 2022. Automatic text summarization of biomedical text data: A systematic review. *Information*, 13(8):393.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

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

John Concato, Nirav Shah, and Ralph I Horwitz. 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine*, 342(25):1887–1892.

Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

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

Jay DeYoung, Eric Lehman, Ben Nye, Iain J Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. *arXiv preprint arXiv:2104.04302*.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Georgios Katsimpras and Georgios Paliouras. 2022. Predicting intervention approval in clinical trials through multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1947–1957, Dublin, Ireland. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, 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.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

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

Irene Li, Alexander Fabbri, Rina Kawamura, Yixin Liu, Xiangru Tang, Jaesung Tae, Chang Shen, Sally Ma, Tomoe Mizutani, and Dragomir Radev. 2021. Surfer100: Generating surveys from web resources on wikipedia-style. *arXiv preprint arXiv:2112.06377*.

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

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

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.

Marcia L Meldrum. 2000. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/oncology clinics of North America*, 14(4):745–760.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. 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*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Julia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022a. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111.

Yulia Otmakhova, Thinh Hung Truong, Timothy Baldwin, Trevor Cohn, Karin Verspoor, and Jey Han Lau. 2022b. LED down the rabbit hole: exploring the potential of global attention for biomedical multi-document summarisation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 181–187, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

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

Sanjana Ramprasad, Denis Jered McInerney, Iain J Marshal, and Byron C Wallace. 2023. Automatically summarizing evidence from clinical trials: A prototype highlighting current challenges. *arXiv preprint arXiv:2303.05392*.

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.

Alexander Te-Wei Shieh, Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2019a. Towards understanding of medical randomized controlled trials by conclusion generation. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 108–117, Hong Kong. Association for Computational Linguistics.

Alexander Te-Wei Shieh, Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2019b. Towards understanding of medical randomized controlled trials by conclusion generation. *arXiv preprint arXiv:1910.01462*.

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.

Arvind Krishna Sridhar and Erik Visser. 2022. Improved beam search for hallucination mitigation in abstractive summarization. *arXiv preprint arXiv:2212.02712*.

Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022a. Investigating crowdsourcing protocols for evaluating the factual consistency of summaries. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5680–5692, Seattle, United States. Association for Computational Linguistics.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022b. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. 2023. Gersteinlab at mediqa-chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. *arXiv preprint arXiv:2305.05001*.

Rahul Tangsali, Aditya Jagdish Vyawahare, Aditya Vyankatesh Mandke, Onkar Rupesh Litake, and Dipali Dattatray Kadam. 2022. Abstractive approaches to multidocument summarization of medical literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 199–203.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

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

David Wan and Mohit Bansal. 2022a. Evaluating and improving factuality in multimodal abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wan and Mohit Bansal. 2022b. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

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

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Tianshu Wang, Faisal Ladhak, Esin Durmus, and He He. 2022. Improving faithfulness by augmenting negative summaries from fake documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11913–11921, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, 252:109460.

Qianqian Xie and Fei Wang. 2023. Faithful ai in healthcare and medicine. *medRxiv*, pages 2023–04.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Shiyue Zhang, David Wan, and Mohit Bansal. 2022b. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. *arXiv preprint arXiv:2209.03549*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.