Scientific Reports

https://doi.org/10.1038/s41598-025-31075-1

Article in Press

A hallucination detection and mitigation framework for faithful text summarization using LLMs

Received: 27 May 2025

Accepted: 28 November 2025

Published online: 03 December 2025

Cite this article as: Liu S., Gao Y., Li S. *et al.* A hallucination detection and mitigation framework for faithful text summarization using LLMs. *Sci Rep* (2025). https://doi.org/10.1038/s41598-025-31075-1

Shenling Liu, Yang Gao, ShaSha Li, PanCheng Wang & Ting Wang

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

A Hallucination Detection and Mitigation Framework for Faithful Text Summarization Using LLMs

Shenling Liu[1*], Yang Gao[2*], ShaSha Li[2], PanCheng Wang[2], Ting Wang[2]

[1*]Education School, National University of Defense Technology, Deyalu Street, ChangSha, 410073, HuNan Province, China.
[2]Computer School, National University of Defense Technology, Deyalu Street, ChangSha, 410073, HuNan Province, China.

*Corresponding author(s). E-mail(s): liushenling@nudt.edu.cn; gy@nudt.edu.cn;
Contributing authors: shashali@nudt.edu.cn; wangpancheng13@nudt.edu.cn; tingwang@nudt.edu.cn;

Abstract

Recent advancements in large language models (LLMs) have significantly propelled the field of automatic text summarization. Nevertheless, the domain continues to face substantial challenges, particularly the issue of hallucination where summaries contain information not present in the source text. Such problems not only compromise the factual accuracy of summaries but also lead to diminished user satisfaction. Existing methods exhibit limitations in effectively detecting and mitigating hallucinations, often lacking transparency in their underlying mechanisms. This paper introduces a hallucination detection and mitigation framework that employs a Question-Answer Generation, Sorting, and Evaluation (Q-S-E) methodology to enable the quantitative detection of hallucinations in summaries. Leveraging LLMs, the framework incorporates an iterative hallucination resolution mechanism, which enhances the transparency of the modification process and improves the faithfulness of text summarization. Experimental results on three benchmark datasets CNN/Daily Mail, PubMed, and ArXiv demonstrate that our approach markedly improves the factual consistency of summaries while preserving their informational completeness.

Keywords: Text Summarization, Hallucination Detection, Q-S-E Methodology, Factual Accuracy

1 Introduction

Text summarization condenses textual material to retain the most essential information, thereby aiding individuals in saving time and effort. With the advancement of end-to-end deep learning technologies, automatic summarization has progressed rapidly. These summarization technologies employ neural networks and pretrained language models to generate coherent summaries. However, these summarization systems encounter numerous challenges in practical applications, with the most significant impact on summary quality arising from the issue of factual consistency, commonly referred to as the hallucination problem. In text summarization, hallucination is defined as the inclusion of any span in the generated summary that is not supported by the input document, rendering it hallucinatory [2]. Previous research indicates that 25% of summaries from the CNN/Daily Mail dataset, generated by traditional end-to-end pretrained language models, exhibit hallucination problems [3]. This reduces the performance of summarization systems and causes the generated summaries to fall short of user expectations in realworld scenarios. The causes of hallucination are multifaceted. In text summarization tasks, hallucination primarily arises from several factors, including imperfect representation learning in the encoder, erroneous decoding by the decoder, and exposure bias during the inference process [1]. These issues result in the generation of content that deviates from the source text, undermining the factual reliability and trustworthiness of summaries. Ji [2] provided a comprehensive overview of the research progress and challenges in addressing hallucination issues in natural language generation. They conducted an indepth analysis of existing studies on hallucination phenomena in various natural language generation tasks such as summarization, dialogue generation, generative question answering, data-to-text generation, machine translation, and visual language generation.Lango et al.[4] explored a novel approach to mitigate hallucinations by combining the probability output of a generator language model with the output of a special 'text critic' classifier. This classifier guides the generation by evaluating the match between input data and text.Zhu et al. [5] proposed injecting a fact-based knowledge graph into the model's decoder, aiming to enhance factual consistency by providing structured, reliable context during the decoding process. While these approaches represents a significant step forward, it relies on parameterized mechanisms that lack transparency. Specifically, researchers and users are unable to pinpoint which parts of the summary are hallucinated or understand the specific modifications made to resolve these inaccuracies. This opacity limits the interpretability of the summarization process and the broader adoption of such methods in domains where traceability and reliability are essential. To address these limitations, this paper proposes a hybrid framework that integrates structured knowledge with data-driven approaches, supporting an iterative correction mechanism. This framework establishes quantitative measures for hallucination detection and leverages the iterative optimization capabilities of large language models (LLMs) to not only enhance the factual consistency of summaries effectively but also improve their interpretability and reliability. Ultimately, the proposed framework seeks to bridge the gap between generation accuracy and user satisfaction, offering a robust solution to the pervasive issue of hallucinations in text summarization. The subsequent sections will detail the methodology employed in this framework, present experimental results validating its effectiveness, and discuss future directions for research in this critical area.

2 Related Work

2.1 Hallucination Detection

Hallucination detection in text summarization primarily encompasses two categories of methods: statistical metric detection and model-based detection. Statistical metrics utilize statistical learning approaches to quantify the presence of hallucination. Building upon methods such as ROUGE and BLEU, Wang et al. [6] proposed PARENT-T, which assesses the fidelity of content by analyzing the overlap and contradictions between generated and reference texts. Dhingra et al.[7]further advanced this concept with PARENT2, incorporating the source text as a reference. Other methods, such as Knowledge F1 and BVSS[8], focus on evaluating different aspects of information quantity. Model-based detection employs neural networks to handle more complex linguistic features, including information extraction, question-answering, and natural language inference technologies, to identify hallucination by validating the consistency of knowledge between generated and reference texts. Among these, the question-answering approach has been used to assess hallucination in various tasks, including summarization [5], dialogue [4], and data-to-text generation[9]. Zhou [10], Santanam et al. [11], and Khot et al.[12] have constructed task-specific datasets to refine metrics based on natural language inference (NLI).

2.2 Hallucination Mitigation

Research on hallucination mitigation primarily focuses on strategies during the training and inference phases. During the pretraining phase, researchers emphasize optimizing the quality of pretraining corpora by reducing noisy data and unverifiable information, thereby lowering the occurrence of hallucinations. Akyürek et al. highlight that a direct approach to reducing hallucinations is carefully curating training data to exclude unverifiable content. Zhou et al. [13] constructed an instruction-tuning dataset during the supervised fine-tuning phase, effectively enhancing the model's fidelity and accuracy. Fernandez et al. [14] integrated reinforcement learning with human feedback (RLHF) to design a specialized reward function aimed at mitigating hallucinations and improving the model's understanding of knowledge boundaries. In the inference phase, Lee et al. [15] proposed a factual nucleus sampling decoding algorithm that combines the strengths of nucleus sampling and greedy decoding to better balance diversity and factuality. Wang et al. [16]adopted chain-of-thought prompting [17], where large language models generate reasoning steps before providing final answers. Additionally, external knowledge injection [18] offers an efficient on-the-fly solution by embedding external knowledge into the generation process, enabling the model to deliver more factual and verifiable responses.

3 Method

To enhance the factual consistency of automatically generated summaries, this chapter introduces a dedicated framework for the detection and mitigation of hallucinations. The proposed framework operates on pre-existing summaries, systematically identifying and rectifying factual inconsistencies to improve their reliability and utility in downstream applications. As illustrated in Figure 1, the framework is composed of two primary modules: the Hallucination Detection Module and the Hallucination Mitigation Module.

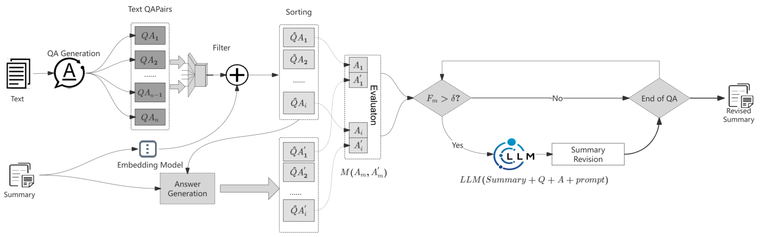

Fig. 1 Model architecture diagram. Our method is divided into the Summary Generation, Hallucination Detection, and Hallucination Mitigation modules.

3.1 Hallucination Detection

The Hallucination Detection Module quantitatively assesses the factual consistency of a generated summary. Central to this module is the proposed G-S-E method, which consists of three sequential stages: (1) Generation of question-answer (QA) pairs from the summary and source document, (2) Sorting of these pairs to prioritize critical evaluations, and (3) Evaluation of factual alignment by comparing answers derived from the summary against those from the source document. This structured approach enables precise identification and quantification of hallucinated content.

3.1.1 QA Generation

In this step, the module generates question-answer (QA) pairs to establish a foundation for evaluating the factual consistency of the summary against the source text. Unlike conventional methods that directly generate QA pairs from the text, our approach first extracts factual statements (i.e., answers) from the source document and then formulates corresponding questions based on these answers. This answer-first QA generation strategy reduces the risk of hallucination in QA pairs, as empirically validated in subsequent experiments. For the sourcetext D, an LLM first identifies key factual statements (answers) using a structured prompt designed to capture critical entities, figures, and events:

> *Extract key factual statements (answers) from the text, focusing on entities, figures, and events.*

Based on these extracted answers, the LLM then generates corresponding questions, forming an initial set of QA pairs:

$$S_D = \{(Q_1, A_1), (Q_2, A_2), \ldots, (Q_n, A_n)\}$$

To ensure quality and relevance, heuristic-based filtering is applied to exclude suboptimal pairs (e.g., overly short, ambiguous, or low-information questions). The refined QA collection is denoted as:

$$S_f = \{(Q_1, A_1), (Q_2, A_2), \ldots, (Q_i, A_i)\}$$

where i represents the number of high-quality QA pairs retained after filtering.

3.1.2 QA Sorting

Following the QA generation and filtering steps, the refined QA pairs are prioritized based on their relevance to the generated summary. To achieve this, we compute the ROUGE-1 F1 score between each question-answer pair (Q_j, A_j) and the summary S. This unigram-based overlap metric serves as a proxy for semantic relevance, enabling the identification of QA pairs that are most closely aligned with the summary's content.

The QA pairs are then ranked in descending order according to their ROUGE-1 F1 scores. The top i most relevant QA pairs are selected and denoted as:

$$S_s = \{(\hat{Q}_1, A_1), (\hat{Q}_2, A_2), \ldots, (\hat{Q}_i, A_i)\}$$

where $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_i$ represent the questions with the highest ROUGE-1 F1 scores relative to the summary S. This subset S_s thus comprises the QA pairs most relevant to the summary, forming the basis for subsequent hallucination evaluation.

3.1.3 Summary-Based Answer Generation

Following the selection and ranking of the top i most relevant QA pairs based on ROUGE-1 scores, the next critical step involves generating corresponding answers directly from the summary S. For each question $\hat{Q}_j$ in the sorted QA set, the model is instructed to extract an answer A'_j from S. This process is illustrated by the following prompt example:

> *Extract information from the above summary to answer: What is the primary cause of death worldwide as of 2020?*

The system generates an answer A'_j from the summary. Repeating this process for all questions in the sorted set results in a new collection of QA pairs that includes answers derived from both the source text and the summary:

$$\text{Source'} \left(\hat{Q}_1, A_1, A'_1\right), \left(\hat{Q}_2, A_2, A'_2\right), \ldots, \left(\hat{Q}_i, A_i, A'_i\right)$$

The newly generated answers $A'_1, A'_2, \ldots, A'_i$ are then used in the subsequent stages to evaluate the factual consistency of the summary.

3.1.4 QA Evaluation

This step systematically quantifies hallucinations within the summary, providing critical data to guide subsequent mitigation strategies. For each question Q, the corresponding answers A (sourced from the original text) and A' (extracted from the summary) are compared. Discrepancies between A and A' reveal factual inconsistencies, highlighting regions where the summary deviates from the source material.

To evaluate the degree of factual consistency between A and A', a scoring function M is employed. The consistency score is formally defined as:

$$\text{Score} = M(A, A')$$

Here, M denotes the evaluation model tasked with measuring the factual alignment between A' and A. In this work, FactCC—a state-of-the-art model tailored for factual consistency assessment—is utilized as the primary evaluation metric. FactCC determines whether the claims in A' are logically and factually entailed by A, enabling fine-grained inconsistency detection.

By applying this evaluation process across all questions in the curated QA set Source^m, we obtain a sequence of consistency scores:

$$\text{F} = \{F_1, F_2, \ldots, F_i\}$$

where F_i represents the consistency score for the i-th QA pair. These scores collectively reflect the degree to which the summary maintains factual fidelity across various aspects of the source content. The overall factual consistency of the summary is then represented by the mean score:

$$F_a = \frac{1}{m}\sum_{i=1}^{m} F_i$$

This aggregate metric serves as a robust indicator of the summary's reliability and factual accuracy.

3.2 Hallucination Mitigation Module

Using the Hallucination Detection Module, the system evaluates whether the hallucination level of a summary surpasses a predefined threshold δ, which is an empirically determined constant, calibrated based on extensive validation experiments conducted on development datasets. This threshold serves as a benchmark for acceptable factual accuracy and consistency. If the detected hallucination level exceeds δ, it signals that the summary is unreliable and requires revision.

To systematically address these inconsistencies, the framework employs prompt engineering a method of crafting tailored inputs to guide the language model toward generating corrected outputs. When a factual inconsistency is identified in a QA pair (Q_k, A_k, A'_k) where A'_k (the answer from the summary) disagrees with A_k (the correct answer from the source) the question Q_k is combined with the accurate answer A_k. A hint is then added to explicitly instruct the model to revise the summary while incorporating the correct information. The query template for this revision is as follows:

> *The key information is wrong or missing; the paper should be revised accordingly, and the following information should be added.*

This structured prompt serves two purposes:

- **Error Localization:** By pinpointing the specific area of inconsistency, the model is focused on the problematic segment of the summary.
- **Guidance for Correction:** The inclusion of the correct answer ensures that the revision is factually grounded in the source text.

Once the revised query is constructed, the LLM generates a new version of the summary:

$$\text{Summary}_r = \text{LLM}(\text{Summary} + Q + A + \text{prompt})$$

The newly generated summary undergoes another round of evaluation by the Hallucination Detection Module to ensure that inconsistencies have been addressed. If hallucinations persist, additional iterations are performed, with the process cycling through evaluation, prompt refinement, and summary revision until the hallucinations are reduced below the threshold δ.

When the revised summary passes the final evaluation, demonstrating factual accuracy and consistency with the input document, it is deemed acceptable. The framework then outputs this refined version as the final summary.

3.3 Ethical Statement

This study, involving human participants, was reviewed and approved by the Research Ethics Committee of the College of Computer Science, National University of Defense Technology. The research was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants prior to their inclusion in the study.

4 Experiments

In this section, we will validate our method on three baseline summary datasets. We use automated metrics and manual evaluations to assess the quality of the generated summaries. We instantiate LLM using ChatGPT.

4.1 Experiment Settings

4.1.1 Datasets

We conducted experiments on the following three publicly available datasets.

CNN/Daily Mail [20] This dataset consists of over 300,000 news articles in the English language written by journalists from CNN and the Daily Mail. Each article in the dataset has a manually written summary.

PubMed[21] The PubMed dataset comprises summaries and articles from scientific and technical literature, primarily sourced from the PubMed database. The content covers domains such as medicine, biology, and chemistry.
Arxiv[22] is a collection designed for the task of summarization of long documents, specifically scientific papers.

4.1.2 Evaluation Metrics

ROUGE[23] is utilized as an automated evaluation metric for assessing the quality of summaries.
BertScore [16] is a metric for measuring text similarity based on the BERT model. It evaluates similarity by calculating the cosine similarity between the embedding codes of two sentences in the BERT model.
FactCC [3] is a weakly supervised, BERT-based model metric that validates factual consistency by applying rule-based transformations to sentences in the source document. It demonstrates a high correlation with human judgment in evaluating summary faithfulness.
BartScore [25] is an evaluation metric for assessing generated text. BartScore evaluates informativeness and factuality.

4.1.3 Baselines

BART [17] An summary pre-training model that disrupts the original text with various types of noise during the pre-training phase, and then reconstructs the original text using a seq2seq model.
GPT-3 [28] developed by OpenAI, is a natural language processing (NLP) model based on the Transformer architecture. It is one of the largest pre-training models to date, featuring 175 billion parameters.
ChatGPT is a conversational model introduced by OpenAI, based on the Generative Pre-trained Transformer (GPT) architecture. It is a specific applica- tion of the GPT-3 model, designed for dialogue generation and understanding.
Ele-aware [19] introduces a new test set that addresses the inconsistency between human preferences for zero-shot summaries in LLMs (Large Language Models) and automatic evaluation metrics.
SummIt [30] achieves iterative improvements in generating summaries through self-assessment and feedback.
Pegasus [27] employs a pre-training objective specifically designed for abstractive summarization, treating gaps in the input document as pseudo-summaries to be generated.

4.2 Implementation

In our framework, ChatGPT (gpt-3.5-turbo) serves as the core large language model (LLM) and is deployed in two key stages: QA pair generation and summary refinement. During the QA pair generation phase, the LLM constructs question-answer pairs based on the source document. In the summary refinement phase, it revises

and improves the generated summary by leveraging QA-based feedback. To ensure reproducibility and consistency, the temperature parameter is set to 0 during all generation and inference processes.
For comparative analysis, we utilize the BART model from Hugging Face and the official checkpoints of PEGASUS as baselines. Additionally, we include Ele-aware and SummIt as reference baselines, both implemented using their publicly available codebases. Specifically, Ele-aware is configured with its default entity-centric settings, while SummIt is executed with its recommended iterative refinement pipeline.
All experiments are conducted on the CNN/Daily Mail dataset cite-hermann2015teaching, using a randomly sampled subset of 1,000 test instances. Hyperparameters and prompt cues for our framework are fine-tuned on a separate development set of 50 examples to ensure optimal performance.

4.3 Result

4.3.1 Quality of Answers Assessment

In the Question and Answer Generation section, we evaluate how effectively the Answer-first QA generation method performs this task and whether the generated answers adequately address the questions, as discussed in this section. Upon analyzing the data samples, we observed that when generating answers, the LLM often extracts sentences directly from the original text, thereby ensuring a certain level of accuracy in the generated responses. Based on this observation, we designed a prompt template that instructs the LLM to extract key entities and numbers from the original text as answers. This approach reduces randomness in answer generation. Specifically, we first generate answers by extracting these key components, followed by generating questions based on the answers. This method ensures a one-to-one correspondence between the questions and answers. We assume that a generated answer is correct if it aligns with the original text. To test this hypothesis, we extracted 50 sets of data and retrieved their corresponding sentences from the original text. As shown in Table 1, the ROUGE-L score is 78.12, indicating that 78.12% of the longest common subsequence in the generated answers matches the source text. Additionally, the ROUGE-1 and ROUGE-2 scores both exceed 70%. These results support our hypothesis that the generated answers are primarily derived from the source text, demonstrating that the answers generated using the LLM effectively address their corresponding questions.

Table 1 Assessment of the quality of questions and answers generated by Answer-first QA generation

ROUGE	Recall	Precision	F1
ROUGE-1	51.02	78.13	61.73
ROUGE-2	44.78	71.42	55.04
ROUGE-L	51.02	78.12	61.72

4.3.2 Summary Quality Evaluation

The results for evaluating generic summary quality are presented in Table 2. We use pre-trained language models, including BART, GPT-3, ChatGPT, Ele-aware, and SummIt, as baseline models. For a fair comparison, we evaluate our method against these baseline models in a zero-shot setting. We observe that the ROUGE score of our method is higher than SummIt on the CNN/Daily Mail dataset, although the BERT-based evaluation metric, BERTScore, is lower than that of Ele-aware. Compared with the output of ChatGPT, our method shows consistent improvement in the BERTScore. Similarly, on the ArXiv and PubMed datasets, our method attains the best overall performance, highlighting its strong suitability for scientific summarization tasks where technical accuracy and conciseness are critical.

Table 2 Experimental results for different models and datasets

Dataset	Method	ROUGE-1	ROUGE-2	ROUGE-L	FACTCC	BertScore	BartScore
CNN/Daily Mail	BART	32.83	13.30	29.64	32.19	88.68	-1.86
	GPT3	30.10	8.98	27.51	32.50	87.79	-1.88
	ChatGPT	34.45	13.98	32.84	35.43	88.80	-1.80
	Ele-aware	37.75	15.20	34.25	36.20	**90.05**	-1.72
	SummIt	36.50	13.49	26.76	36.00	89.59	-1.70
	Ours	**37.80**	**15.88**	**35.77**	**36.61**	89.91	**-1.70**
Pubmed	GPT3	29.30	10.98	27.33	32.50	85.18	-1.91
	ChatGPT	30.16	11.04	28.15	35.10	86.05	-1.88
	Ours	**30.98**	**11.33**	**28.75**	**37.78**	**88.14**	**-1.76**
Arxiv	GPT3	30.10	11.03	28.57	32.37	84.34	-1.98
	ChatGPT	32.43	11.12	30.16	36.00	85.31	-1.83
	Ours	**33.29**	**12.08**	**30.92**	**38.14**	**87.02**	**-1.75**

4.3.3 Faithfulness Evaluation

To further assess the effectiveness of our method in maintaining factual consistency, we conducted a series of faithfulness-oriented experiments, with results summarized also in Table 2. As the results indicate, on the CNN/Daily Mail dataset, our method achieves the best scores on FactCC and BartScore, and even improves the FactCC metric by 1.1 points compared to the results of ChatGPT. And on the Pubmed dataset, our iteratively optimized results are 0.12 higher on the Bartscore metric and 2.09 points higher on the FactCC metric compared to the summaries generated by ChatGPT. On Arxiv dataset, our method also achieves the best performance. The findings suggest that our method produces significant improvements in refining summaries. In conclusion, our method effectively enhances the faithfulness of summaries without compromising the information content of a summary.

4.3.4 Human Evaluation

Based on previous findings that human annotators tend to favor summaries generated by the LLM model even when ROUGE scores are relatively low [31], we further validated the effectiveness of SummIt through a specialized human study. Specifically, we used a five-point Likert scale including coherence, fluency, relevance, consistency, conciseness, and overall evaluation of the summaries, the results are shown in Table 3. Based on the five-point Likert scale, summaries generated by our method outperformed the results of the pre-trained language model. Our method also produced improvements in factual consistency, which consistent with the results demonstrated by FactCC. In addition, we observed a significant improvement in human annotators' preference for our summaries after iterative modifications, suggesting that the improved method is effective in generating general summaries.

Table 3 Human Evaluation Results

Method	Coherence	Fluency	Relevance	Consistency	Conciseness	Overall
PEGASUS	3.90	4.00	4.30	3.20	3.00	3.68
BART	3.80	4.10	4.00	3.70	3.70	3.74
ChatGPT	4.20	4.40	4.30	4.10	3.90	4.18
Ours	4.30	4.30	4.50	3.70	4.20	4.20

4.4 Ablation Experiment

In this section, we will discuss the roles played by different modules of our method. Our improvements mainly include the iterative modification part and the question sorting part, and we will discuss the impacts of these two parts on the final results and conduct ablation experiments to validate them. In order to verify the effectiveness of our method, we designed a one-time iterative prompt template as a control group, and we connected all the QA pairs in the QA set together and fed them into the LLM, so that the LLM could complete the modification of the summaries at one time. In order to verify the effectiveness of our question ordering module, we did the corresponding ablation experiments, the results of which are shown in Table 4. We extracted 50 sets of data on Pubmed and CNN/Daily Mail datasets respectively for the experiments. On each dataset, the method based on multiple rounds of iteration and problem ranking performs the best on most of the evaluation metrics, with the highest ROUGE-1, ROUGE-2, ROUGE-L, FACTCC, BertScore and BartScore. This indicates that our method is more effective compared to other methods. The single-word iterative approach usually performs better than the ChatGPT approach, suggesting that performing a single iteration on the model may im- prove its performance. The multiple iteration method usually performs better than the single iteration method, which suggests that multiple iterations improve performance without sorting.

4.5 Impact on Summary Informativeness

As previously demonstrated in our experiments, the proposed framework effectively enhances summary faithfulness through fact-based question-answering (QA) alignment. However, a natural question arises: does this improvement in faithfulness come at the cost of summary informativeness? To investigate this, we conducted a dedicated evaluation of summary informativeness under constrained summary length conditions.

Table 4 Ablation Experiment

Method	ROUGE-1	ROUGE-2	ROUGE-L	FACTCC	BertScore	BartScore
			CNN/Daily Mail			
ChatGPT	34.45	13.98	32.84	35.43	88.80	-1.80
One iteration	36.50	13.49	26.76	36.00	89.59	-1.70
iterations + No sort	37.80	15.88	35.77	36.61	89.91	-1.70
			Pubmed			
ChatGPT	30.16	11.04	28.15	35.10	86.05	-1.88
One iteration	30.98	11.33	28.75	37.78	88.14	-1.76
iterations + No sort	30.98	11.33	28.75	37.78	88.14	-1.76

During summary generation, we observed that when the model is optimized solely for faithfulness, some important information from the source document may be omitted. To quantitatively analyze this issue, we designed an experimental protocol based on question generation and response matching. Specifically, questions were generated from the source articles. If the content required to answer a question was absent from the summary, the question was classified as an *other question*; if the summary could answer the question, it was labeled as a *factual question*. Through this process, two distinct question sets were constructed: one for factual questions and one for other questions.

Table 5 Summary Informative Research Result

Method	ROUGE-1	ROUGE-2	ROUGE-L	FACTCC	BertScore	BartScore
			Pubmed			
Other Questions	32.28	12.82	29.10	87.30	87.30	-1.91
Faithful Questions	31.19	27.74	37.55	89.10	89.10	-1.73
Both Questions	32.33	12.21	29.30	37.78	89.10	-1.72

We evaluated the impact of using each question set independently. For this experiment, we extracted 50 document-summary pairs from the PubMed dataset. The experimental results are presented in Table 5. The results show that when only factual questions were used for evaluation, the ROUGE scores of the summaries decreased by approximately 2%. Conversely, when only other questions were used, the faithfulness of the summaries significantly declined. These findings suggest that while augmenting summary faithfulness is beneficial, it may inadvertently lead to the loss of important information, thereby compromising the overall informativeness of the summary.

4.6 Effect of Iteration Number

In this section, we investigate the effect of the number of iterative refinement steps on the quality of generated summaries within our proposed framework. Given that the framework performs multiple rounds of modification to enhance summary faithfulness, a critical question arises: do subsequent refinement steps undermine previously improved aspects of the summary, such as informativeness or factual consistency? To address this concern, we conducted a controlled experiment to explore the influence of iteration count on summary quality. The experiments were conducted on a subset of the PubMed dataset containing 50 examples. Summary informativeness was evaluated using ROUGE and BERTScore, while factual consistency was assessed using FactCC and BartScore. These metrics collectively provide a comprehensive assessment of summary quality across both informational and factual dimensions. The experimental results reveal that both ROUGE and BERTScore scores initially increase and then decrease as the number of iterations grows. This trend indicates that iterative refinement can indeed enhance the overall quality of summaries; however, excessive iterations may lead to a loss of critical information. Similarly, the factual consistency metrics, FactCC and BartScore, exhibit the same trend, reinforcing the conclusion that simply increasing the number of refinement steps does not indefinitely improve summary quality. Therefore, it is essential to control the number of iterative modifications. Based on our experimental findings, we recommend limiting the refinement process to approximately six iterations to achieve the optimal trade-off between factual accuracy and information retention.

5 Conclusion

This paper introduces a framework that effectively mitigates hallucinations in text summarization. Our method consists of a hallucination detection component and an iterative modification component, instantiated using a large language model. We conducted experiments and analyses on three benchmark datasets, and the results demonstrate that summaries generated by our method outperform those from a single pass of an LLM, validating the effectiveness of our approach. While our current evaluation focuses on factual consistency, future work could extend this framework to time-sensitive domains. This would involve assessing metrics such as Carburacy to ensure the summaries are not only faithful but also temporally accurate. We hope that our method and this identified research direction can guide future efforts in constructing more robust and reliable LLM-based summarization systems.

6 Data Availability Statement

The research presented in this manuscript primarily utilized data from three distinct public sources: CNN/Daily Mail dataset: This dataset, commonly used for summarization tasks, contains news articles and their corresponding summaries. It is publicly accessible via platforms such as Hugging Face Datasets (https://huggingface.co/google/pegasus-cnn_dailymail). PubMed database: A comprehensive public archive of biomedical and life sciences journal literature, hosted by the National Library of Medicine (NLM) at the National Institutes of Health (NIH). It can be freely accessed online at https://pubmed.ncbi.nlm.nih.gov/. arXiv e-Print archive: A highly regarded open-access repository of electronic preprints (known as e-prints) approved for publication after moderation, covering areas in physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering, and economics. It is publicly available at https://arxiv.org/. All data used in this study were sourced directly from these publicly available platforms. No new datasets were generated as part of this research.

References

[1] Rehman T , Mandal R , Agarwal A ,et al.Hallucination Reduction in Long Input Text Summarization[J]. 2023.

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” ACM Computing Surveys, vol. 55, pp. 1-38, 2023.

[3] W. Kryściński, B. McCann, C. Xiong, and R. Socher, “Evaluating the factual consistency of abstractive text summarization,” ArXiv Preprint ArXiv:1910.12840, 2019.

[4] M. Lango and O. Dušek, “Critic-driven decoding for mitigating hallucinations in data-to-text generation,” ArXiv Preprint ArXiv:2310.16964, 2023.

[5] C. Zhu, W. Hinthorn, R. Xu, et al., “Boosting factual correctness of abstractive summarization with knowledge graph,” ArXiv Preprint ArXiv:2003.08612, 2020.

[6] L. Wang, et al., “PARENT-T: A new metric for generative models,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 400–409.

[7] B. Dhingra, et al., “PARENT: A new metric for generative models,” arXiv preprint arXiv:1801.09839, 2018.

[8] M. Martindale, et al., “BVSS: A new metric for NMT adequacy,” arXiv preprint arXiv:1804.08771, 2018.

[9] T. Khot, M. Sap, A. Sabharwal, and P. Clark, “SciTaiL: A textual entailment dataset from science question answering,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[10] Y. Zhou, T. Wu, Y. Jiang, et al., “DeepNup: Prediction of Nucleosome Positioning from DNA Sequences Using Deep Neural Network,” Genes, vol. 13, no. 11, pp. 36360220, 2022.

[11] Y. Santhanam, S. Yoganathan, V. Sivakumar, et al., “Predictors of Outcome in Children with Status Epilepticus during Resuscitation in Pediatric Emergency Department: A Retrospective Observational Study,” Annals of Indian Academy of Neurology, vol. 20, pp. 142–148, 2017.

[12] O. Honovich, T. Scialom, O. Levy, et al., “Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor,” in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 14409–14428.

[13] C. Zhou, P. Liu, P. Xu, et al., “Lima: Less is more for alignment,” Advances in Neural Information Processing Systems, vol. 36, 2024.

[14] P. Fernandez, D. Garcia de la Garza, and J. Fernández Acín, “Survey: Market Risk Premium and Risk-Free Rate used for 80 countries in 2023,” SSRN Electronic Journal, 2023.

[15] J. P. Lee, C. Binger, N. Harrington, et al., “Aided language measures: Establishing observer agreement for communicators in early language phases,” American Journal of Speech-Language Pathology, 2022.

[16] Y. Wang, Z. Zhang, and R. Wang, “Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method,” ArXiv Preprint ArXiv:2305.13412, 2023.

[17] J. Wei, X. Wang, D. Schuurmans, et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” arXiv preprint arXiv:2201.11903, 2022.

[18] J. Ren, et al., “Using external knowledge as supplementary evidence to assist LLMs in providing truthful responses,” SGD - Saccharomyces Genome Database, 2023.

[19] Y. Wang, Z. Zhang, and R. Wang, “Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method,” ArXiv Preprint ArXiv:2305.13412, 2023.

[20] K. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, “Teaching machines to read and comprehend,” Advances In Neural Information Processing Systems, vol. 28, 2015.

[21] F. Dernoncourt and J. Lee, “Pubmed 200k rct: a dataset for sequential sentence classification in medical summaries,” ArXiv Preprint ArXiv:1710.06071, 2017.

[22] B. Hu, Q. Chen, and F. Zhu, “LCSTS: A Large Scale Chinese Short Text Summarization Dataset,” Proceedings Of The 2015 Conference On Empirical Methods In Natural Language Processing, pp. 1967-1972, 2015.

[23] C. Lin, “Rouge: A package for automatic evaluation of summaries,” Text Summarization Branches Out, pp. 74-81, 2004.

[24] T. Zhang, V. Kishore, F. Wu, K. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” ArXiv Preprint ArXiv:1904.09675, 2019.

[25] W. Yuan, G. Neubig, and P. Liu, “Bartscore: Evaluating generated text as text generation,” Advances In Neural Information Processing Systems, vol. 34, pp. 27263-27277, 2021.

[26] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” ArXiv Preprint ArXiv:1910.13461, 2019.

[27] Alsultan, R., Sagheer, A., Hamdoun, H., Alshamlan, L.,Alfadhli, L. (2025). PEGASUS-XL with saliency-guided scoring and long-input encoding for multi-document abstractive summarization. Scientific Reports, 15(1), 26529.

[28] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and Others, “Language models are few-shot learners,” Advances In Neural Information Processing Systems, vol. 33, pp. 1877-1901, 2020.

[29] Y. Wang, Z. Zhang, and R. Wang, “Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method,” ArXiv Preprint ArXiv:2305.13412, 2023.

[30] H. Zhang, X. Liu, and J. Zhang, “SummIt: Iterative Text Summarization via ChatGPT,” ArXiv Preprint ArXiv:2305.14835, 2023.

[31] T. Goyal, J. Li, and G. Durrett, “News summarization and evaluation in the era of gpt-3,” ArXiv Preprint ArXiv:2209.12356, 2022.