Citation-Enhanced Generation for LLM-based Chatbot

Anonymous ACL submission

Abstract

Large language models (LLMs) exhibit powerful general intelligence across diverse scenarios, including their integration into chatbots. However, a vital challenge of LLM-based chatbots is that they may produce hallucinated content in responses, which significantly limits their applicability. Various efforts have been made to alleviate hallucination, such as retrieval augmented generation and reinforcement learning with human feedback, but most of them require additional training and data annotation. In this paper, we propose a novel post-hoc **C**itation-**E**nhanced **G**eneration (**CEG**) approach combined with retrieval argumentation. Unlike previous studies that focus on preventing hallucinations during generation, our method addresses this issue in a post-hoc way. It incorporates a retrieval module to search for supporting documents relevant to the generated content, and employs a natural language inference-based citation generation module. Once the statements in the generated content lack of reference, our model can regenerate responses until all statements are supported by citations. Note that our method is a training-free plug-and-play plugin that is capable of various LLMs. Experiments on various hallucination-related datasets show our framework outperforms state-of-the-art methods in both hallucination detection and response regeneration on three benchmarks. Our codes and dataset will be publicly available.

1 Introduction

Large Language Models (LLMs) have experienced rapid development in recent years, which show powerful general intelligence in various scenarios (Yue et al., 2023; Singhal et al., 2023). Current LLM-based chatbots, epitomized by ChatGPT and GPT-4, demonstrate impressive capabilities across distinct domains in communicating with humans. There is a growing consensus that LLM-based chatbots can be the next generation of information acquisition methodology.

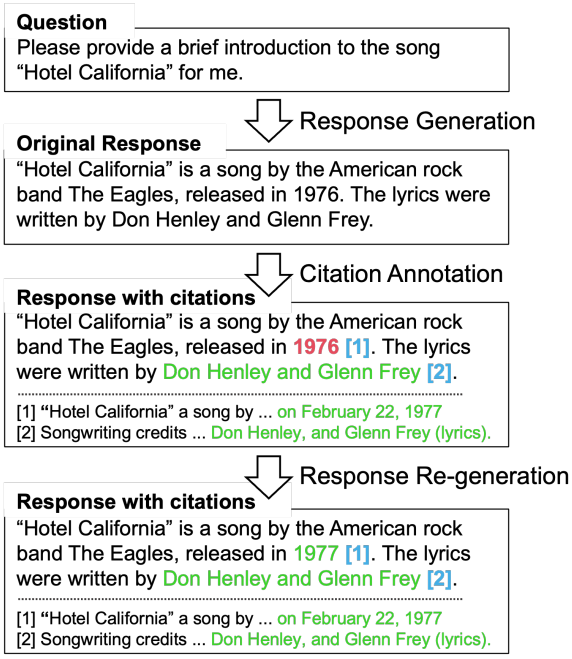

Figure 1: An illustration of our method, which adds citations for the generated content. If there are hallucinations in the generated content, we prompt LLM to regenerate a new response.

However, a critical and unsolved challenge of LLM-based chatbots is the hallucination problem (Ji et al., 2023), which indicates these chatbots may generate hallucinated content in responses randomly. As the underlying mechanisms of hallucinations remain unclear, this problem has substantial constraints on the deployment of LLM-based chatbots in various sensitive scenarios, such as healthcare and education, where reliability is paramount.

Previous approaches have attempted to mitigate this issue through retrieval augmentation (Borgeaud et al., 2022; Izacard et al., 2022) and value alignment (RLHF) (Ouyang et al., 2022; Touvron et al., 2023) in response generation, but these often require additional training and extensive data annotation. For example, InstructGPT (Ouyang et al., 2022) utilize RLHF to alleviate hallucina-

tions in model output, but needs extra training. Gao et al. (2023a) attempt to reduce hallucination through adding retrieved related documents and citations before generation, while the pre-hoc way of incorporating citations may potentially harm the model performance, resulting in poor response results with hallucinations.

In this work, we propose a novel method to alleviate hallucination in LLMs, which leverages retrieval augmentation and Natural Language Inference (NLI) technologies to implement Citation-Enhanced Generation (CEG) in a post-hoc way. Figure 1 is an illustration. Differing from previous studies, the retrieval augmentation module of the CEG framework works after generation (post-hoc), and CEG prompts the model to regenerate the answer when necessary. This approach is effective and easy to use, which can reduce the hallucination in the model's output for various LLMs. We conduct experiments on distinct hallucination-related benchmarks, including detection and response regeneration. where our method achieved state-of-the-art performance. Further analyses demonstrate the usefulness of each module on CEG.

In summary, the main contributions of our work can be summarized as follows:

- We are the first to propose the use of citation to alleviate hallucination in a post-hoc way with regeneration.
- We design a novel post-hoc citation-enhanced generation framework combined with retrieval augmentation and NLI to avoid hallucinations, which is flexible for existing LLMs.
- Experimental results show that our CEG framework achieves the best performance on three hallucination-related benchmarks.

2 Related Work

2.1 Hallucination Control in LLMs

Generative AI has achieved significant advancements, while still facing the hallucination problem. Existing strategies can be categorized into major two types: mitigation during training and mitigation during inference. For the first type, LLMs, such as LLaMA 2 (Touvron et al., 2023), undergo extensive training cycles with high-fidelity data sources like Wikipedia to bolster factual consistency in pre-training. Zhou et al. (2023) alleviate hallucination during instruction fine-tuning, which adopts high quality manually annotated content to regulate hallucination. Some studies (Ouyang et al., 2022; Touvron et al., 2023) also introduce penalties for nonfactual responses to alleviate hallucination in RLHF. However, all these methods need extra training and annotations.

On the other hand, researchers try to deal with the hallucination challenge during inference. Inference-Time-Intervention (Li et al., 2023b) mitigates hallucination by shifting model activations along these factuality-related directions during inference. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has become a prevalent technique in alleviating hallucination by retrieving reliable documents before generation (Yu, 2022). While these methods still generate hallucinations due to the lack of post-hoc verification and they are unable to provide citations for verification.

2.2 Citation Augmented LLMs

In the realm of LLMs, retrieval technology has become a crucial component (Zhang et al., 2023b; Gao et al., 2023b), as it provides related knowledge in generating more reliable results (also mitigates the occurrence of hallucinations). Previous studies point out that citation, generated by retrieval models, is key to building responsible and accountable LLMs (Huang and Chang, 2023).

Existing citation augmented strategies can be divided into two types: parametric and non-parametric. Parametric methods (Taylor et al., 2022) refer to information internalized from the training data, often leading to inaccurate annotated documents, as the annotation process itself can give rise to hallucinations. Non-parametric methods (Gao et al., 2023a; Menick et al., 2022; Izacard and Grave, 2021) involve querying relevant information and seamlessly integrating the retrieved content from outside corpus, which provides more reliable citations. Thus, most previous studies are non-parametric, but they are pre-hoc based. For example, Gao et al. (2023a) adopt retrieval processes to facilitate the annotation of documents within model-generated outputs. Nevertheless, their pre-hoc annotation strategy inadvertently escalates the complexity of a QA task by converting it into a dual challenge of generating a response coupled with simultaneous annotation. Different from existing citation augmented studies, we propose a different strategy to utilize retrieval models to generate citations in a post-hoc way.

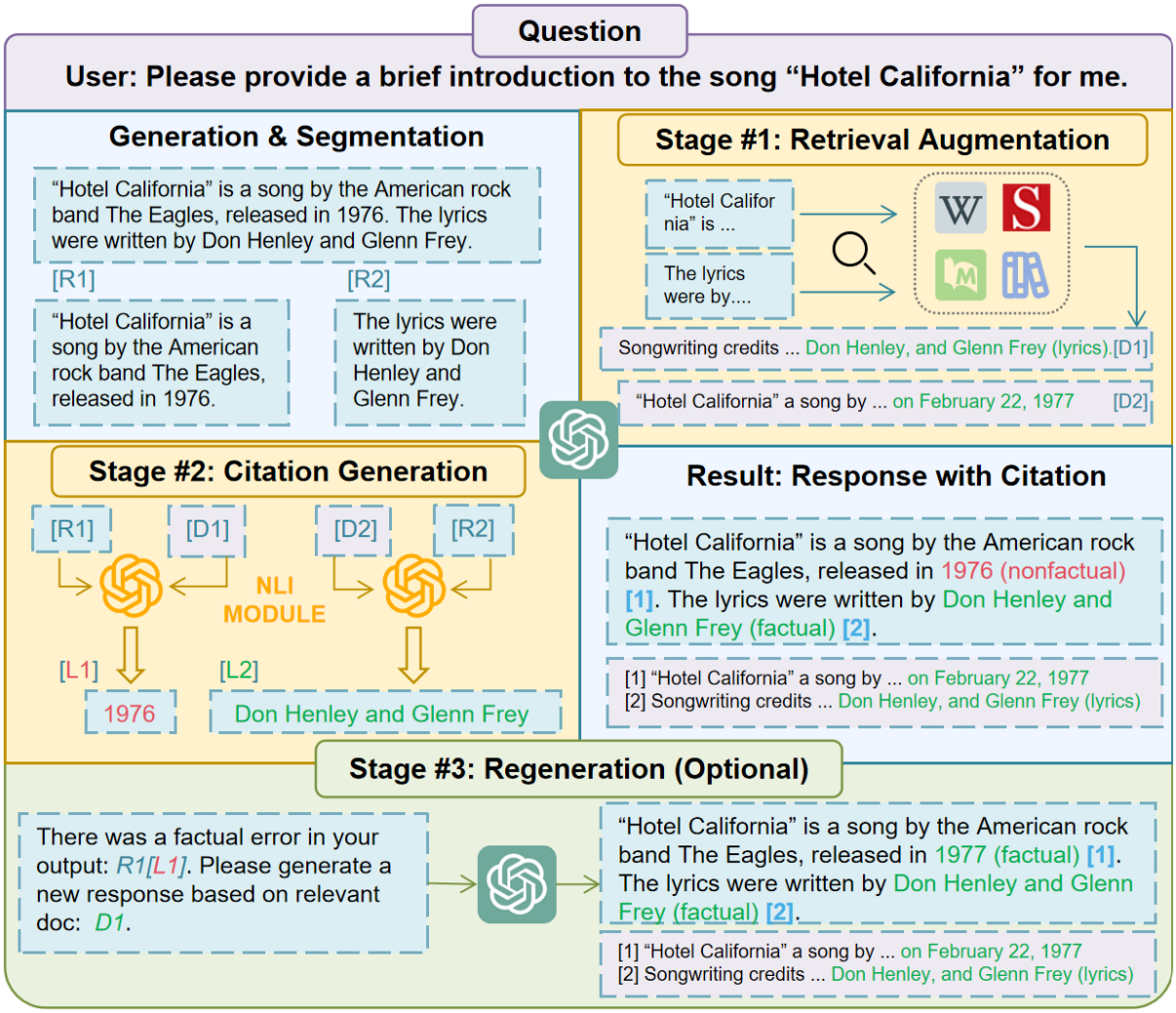

Figure 2: An overview of our CEG framework. [R1] and [R2] denote segments. [D1] and [D2] represent retrieved documents for each segment. [L1] and [L2] are labels (Factual/Nonfactual) generated by the NLI module.

3 Approach

3.1 Overview

Firstly, we give an overview of our proposed CEG framework. Illustrated in Figure 2, CEG has several critical modules: 1) Retrieval augmentation module, designed to search for documents D_j relevant to the original response R. In cases where responses are excessively lengthy, they can be broken down into sub-claims $R = \{R_1, R_2, ..., R_n\}$. 2) Citation generation module, which assesses if the retrieved documents D_j substantiate the $\{R_i\}$ in response or not. 3) Regeneration module, tasked with creating a new prompt that integrates the original user query and key retrieved information for the LLM M to get a more reliable response R'.

It is important to note that our method is a post-hoc framework and is highly adaptable across different LLMs, as it does not require any additional training or fine-tuning. Consequently, we do not specify a particular LLM here.

3.2 Retrieval Augmentation Module

Retrieval augmentation has been shown to have powerful abilities in previous hallucination-related studies (Gao et al., 2023a; Zhao et al., 2023). Different from these studies that aim to retrieve documents as evidence before response generation (questions are queries), we propose to conduct retrieval augmentation in a post-hoc way to verify the correctness of the generated claim R_i (claims are queries). As there are various existing studies on how to retrieve the most related document, we use a simple but effective dense retrieval strategy to verify the performance of our CEG framework, and we believe stronger retrieval will bring further improvements.

Query: For the response R, it will be seg-

mented into several claims if necessary, resulting in $R = \{R_1, R_2, ..., R_n\}$. Here, we adhere to previous work (Chen et al., 2023) and employ a heuristic algorithm for segmentation. Utilizing the NLTK (Bird et al., 2009) sentence tokenizer, we split R to obtain reasonable results that align with user reading habits to get the claims R_i. Then, R_i is adopted as the query one by one.

Corpus (Candidate Documents): The choice of Corpus decides the scope of applications, and there are multiple candidates. In this study, our focus lies predominantly in the domain of knowledge-based question answering, necessitating the employment of a curated corpus. To this end, we leverage a processed snapshot of Wikipedia from October 20, 2023[1], segmented into approximately 100-word candidate documents, each demarcated by a period or newline character. Note that you can replace it with any other corpus, and we use it as most hallucination benchmarks are based on Wikipedia.

Retriever: Dense vector based retrieval technologies have demonstrated powerful performances in recent years, which are also widely used in existing RAG models. Here, we adopt the SimCSE BERT[2] (Gao et al., 2021; Wang et al., 2023d) as the query and document encoder for its promising efficiency in previous studies (Wang et al., 2023c,d). Then, candidate documents are ranked based on cosine similarity scores calculated by the following equation:

$$\text{Sim}(R_i, d_j) = \frac{e(R_i) \cdot e(d_j)}{\|e(R_i)\| \cdot \|e(d_j)\|},$$

Where $e()$ is the SimCSE BERT encoder, d_j is a candidate document in the corpus. As more documents need more calculation in further modules, the top-k retrieved documents with higher similarity are selected to form the reference document set $D_i = \{d_{i,1}, d_{i,2}, ..., d_{i,k}\}$. We add an extra threshold t to filter out the retrieved documents that have low cosine similarity. Apart from the top-1 document, if the $\text{Sim}(R_i, d_j) < t$, d_j will not be included in D_i. These documents are subsequently concatenated to construct the final retrieved content D_i for further calculation.

3.3 Citation Generation Module

After getting the reference document D_i for each response segment R_i, the next step involves generating labels and citations to verify the correctness of R_i. We propose to adopt an NLI method to determine the relationship between each claim-document pair (R_i, D_i). In general, the relationship can be categorized into three types: support, independence, and contradiction. But in hallucination-related scenarios, to adhere to previous studies, we only utilize two categories: (1) **Factual**, where D_i serves as a reference for R_i, thereby substantiating the claim. (2) **Nonfactual**, which means D_i presents a opposite claim to R_i.

Although numerous models (Honovich et al., 2022; Raffel et al., 2020) are capable of the NLI method, our CEG framework seeks to fully leverage the language comprehension capabilities of LLMs. Therefore, we prefer to utilize an LLM, equipped with predefined prompts, to serve as the NLI method. An illustrative prompt is provided below:

- -

Instruction: I will show you a question, a response segment of this question, and a reference document. Your task is to assess whether the given response segment contains factual errors or not with the help of the reference document. ...

- -

When the LLM output indicates "factual", the document D_i is identified as a valid reference for the claim R_i. Consequently, this citation can be added to the original response. If none of the retrieved top-k documents substantiate the claim R_i or if there are documents opposing the claim, we will label this claim as nonfactual (potential hallucination) to remind users to keep carefully reading. Based on the introduced two modules, we can detect whether there are hallucinations in responses.

3.4 Response Regeneration Module

In previous modules, our framework offers a post-hoc method to conduct citation-enhanced verification for responses, where reliable responses are incorporated with citations. However, a new challenge is how to deal with potential hallucinations. So we propose a response regeneration module.

Assuming LLM M generates the original response R, our framework will provide a new prompt for regeneration. The prompt not only contains the original query, but is also incorporated with retrieved documents and the annotated nonfactual segments. Here we provide a brief illustration of the prompt (a full prompt is shown in appendix):

[1] https://dumps.wikimedia.org/enwiki/

[2] https://huggingface.co/princeton-nlp/sup-simcse-bert-base-uncased/tree/main

- -

User: Question; **Chatbot**: Response; **User**: [There were some factual errors in your output: (*Nonfactual Claims*). Please generate a new response based on relevant docs: (*Relevant Docs*).]

- -

Upon receipt of the regenerated response, we can initiate a new citation-enhanced generation process. If the response is adjudged to be free of factual errors, it becomes the final response and will be shown to users. However, if the new response still contains hallucinations, the regeneration cycle will be repeated. To conserve API resources and reduce the waiting time, a predefined parameter T can be set to constrain the max regeneration attempts.

4 Experimental Settings

4.1 Overview

To verify the effectiveness of our framework, we adopt four hallucination-related datasets in our experiments: WikiBio GPT-3 (Manakul et al., 2023), FELM (Chen et al., 2023), and HaluEval (Li et al., 2023a), and WikiRetr. WikiBio GPT-3 and FELM are hallucination detection benchmarks. HaluEval is a hallucination generation benchmark. Besides, we construct a new dataset named WikiRetr, which is to evaluate the retrieval and citation annotation performance. Due to the tasks and baselines are distinct in various datasets, we will introduce each dataset and corresponding settings one by one.

We use GPT models as the LLM backbones, and the version involved in different datasets is distinct for fair comparison with existing baselines. Unless otherwise specified, "ChatGPT" refers to GPT-3.5-Turbo-1106, and "GPT-4" refers to GPT-4-0613. We set the decoding temperature as 0 to maintain the reproducibility of the responses generated by LLMs. All prompts are listed in Appendix A, and more dataset information is shown in Appendix B.

4.2 WikiBio GPT-3 Dataset

WikiBio GPT-3 dataset is constructed to evaluate the hallucination of LLMs. Researchers randomly select 238 biographical articles from WikiBio dataset (Lebret et al., 2016), and utilize the text-davinci-003 to generate new passages. The passages are split into 1,908 sentences, and then manually annotated into three categories: *Major Inaccurate*, *Minor Inaccurate*, and *Accurate*. Following previous studies, *Major Inaccurate* and *Minor Inaccurate* are categorized as **Nonfactual** (potentially with hallucinations, 1,392 segments), and *Accurate* is treated as **Factual** (516 segments).

Baselines: 1) HalluDetector (Wang et al., 2023b) utilizes external knowledge sources, a specific classification model and a Naive Bayes classifier to detect hallucination. 2) Focus (Zhang et al., 2023a) adopts a multi-stage decision-making process, where both pre-retrieval and task specific classifiers are adopted. 3) SelfCheckGPT [3], three variants of which are utilized: w/BERTScore, w/Prompt, and w/NLI (Manakul et al., 2023). SelfCheckGPT w/BERTScore is based on the inherent uncertainty of LLM, while SelfCheckGPT w/Prompt and w/NLI draw upon external knowledge sources. The Area Under the Precision-Recall Curve (AUC-PR) and Balanced_Accuracy are adopted as evaluation metrics.

4.3 FELM Dataset

FELM dataset is designed to evaluate hallucination detection ability. Researchers assemble prompts from diverse scenarios, and use them to instruct GPT-3.5-Turbo-0301 to generate responses. These responses are manually annotated as nonfactual and factual, along with supporting documents. Our experiments are conducted on the WorldKnowledge subset of FELM as it is based on Wikipedia corpus.

Baselines: Following settings in FELM, we adopt four strategies with ChatGPT, GPT-4, and Vicuna-33B as the backbone LLM (Zheng et al., 2023): 1) Vanilla prompts. 2) Prompts augmented with Chain-of-Thought (CoT) reasoning (Kojima et al., 2022). 3) Prompts augmented with hyperlinks to reference documents and 4) Prompts augmented with human-annotated reference documents (Chen et al., 2023). Experiments are conducted at the individual response level. The metrics are accuracy of nonfactual, factual, and balanced.

4.4 HaluEval Dataset

HaluEval dataset is a benchmark for assessing the ability of LLMs to discern hallucinations. Each instance comprises a question, a correct answer, and a hallucinated answer (multiple answers are automatically generated, and the most confusing one is selected by ChatGPT). The QA subset of HaluEval is adopted as it is constructed by Wikipedia corpus, and 2,000 samples of which are randomly sampled.

Baselines: We adopt several models building upon the updated version of ChatGPT as previ-

[3]The latest version in https://arxiv.org/pdf/2303.08896.pdf.

Method		AUC-PR (%)		Balanced_Acc (%)
		Nonfactual	Factual	
HalluDetector	$C_M = 14$, $C_{FA} = 24$	82.42	57.01	70.54
	$C_M = 28$, $C_{FA} = 96$	86.45	61.96	74.82
Focus	$LLaMA-30B_{focus}$	89.79	65.69	73.64
	$LLaMA-65B_{focus}$	89.94	64.90	74.08
SelfCheckGPT	w/BERTScore	81.96	44.23	59.31
	w/NLI	**92.50**	58.47	70.55
	w/Prompt	91.16	68.37	72.64
CEG	top-k=6	92.31	**70.24**	**77.59**

Table 1: Experimental results of our method powered by GPT-3.5-Turbo-Instruct on WikiBio GPT-3. The Self-CheckGPT with Prompt is also powered by GPT-3.5-Turbo-Instruct because GPT-3.5-Turbo-0613 is deprecated.

ous studies (Li et al., 2023a): 1) Vanilla prompts. 2) Prompts augmented with CoT reasoning. 3) Prompts with Pre-RAG, where a strong and fine-tuned retriever, All-mpnet-base-v2[4], is used. Accuracy is chosen as the evaluation metric.

4.5 WikiRetr Datasets

WikiRetr datasets are designed to conduct further analyses on our CEG framework, which is created based on the October 20, 2023 snapshot of Wikipedia. We randomly select 1,000 passages, and apply text-davinci-003 and GPT-4 to rewrite them as new claims, separately. So that each rewritten claim is accompanied by an original passage. The constructed datasets are named WikiRetr-GPT3 and WikiRetr-GPT4, respectively.

To analyze the retrieval module, we utilize various retrievers, including: 1) SimCSE BERT, which is employed in our CEG framework; 2) Sentence BERT (Reimers and Gurevych, 2019), a retriever trained with siamese networks; and 3) All-mpnet-base-v2. Recall@k is the metric to verify if the original document is retrieved in top-k. Precision@k is the metric to verify if the claim is supported by a doc in top-k docs with NLI method.

For NLI methods in the citation generation module, we randomly select 100 instances from each dataset for evaluation. We conduct manual annotation to assess whether the original passages can support the rewritten claims by three annotators. Labeling results show that 90% and 94% generated claims are supported by original documents, which is the ground truth of NLI models. Then, we use 1) True-9B[5] (Honovich et al., 2022) model and 2) GPT models as NLI methods in analyses. We choose the consistency between manual and model annotations as the evaluation metric.

5 Experimental Results and Analyses

5.1 Performance on Hallucination Detection

To verify the effectiveness of our method, we utilize our retrieval augmentation and citation generation modules for hallucination detection on WikiBio GPT-3 and FELM datasets.

Overall performances in WikiBio GPT-3 dataset are shown in Table 1, and we have the following observations: 1) Our CEG framework outperforms all baseline methods in Balanced_ACC and AUC-PR of Factual segments, achieving the second performance in AUC-PR of nonfactual segments. These results indicate the strength of our method. 2) Previous pre-retrieval augmented generation models, SelfCheckGPT w/NLI and w/prompt, also get good performances in AUC-PR. While suffering from the lengthy text, they cannot achieve comparable performance of our model in all metrics. 3) Our model performs slightly worse than w/NLI in the AUC-PR of nonfactual segments, the reason can be the NLI module of SeftCheckGPT undergoes additional training on detecting nonfactual segments (but achieve poor results in factual).

Experimental results in FELM dataset are summarized in Table 2. Firstly, our model achieves the best result in balanced accuracy with GPT-4, indicating its effectiveness. Most baseline models are biased in classifying a single type. Then, CEG with ChatGPT beats other ChatGPT baselines, showing the flexibility of our model. Thirdly, CEG outperforms all pre-retrieval baselines, which shows the strength of the proposed post-hoc segment-level retrieval module in hallucination detection.

[4]https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[5]https://huggingface.co/google/t5_xxl_true_nli_mixture.

Method		Accuracy (%)		
		Nonfact	Factual	Balanced
Vanilla	Vicuna-33B	72.8	34.0	53.4
	ChatGPT	3.4	96.1	49.8
	GPT-4	21.8	95.3	58.5
CoT	Vicuna-33B	40.8	62.3	51.6
	ChatGPT	2.7	**96.9**	49.8
	GPT-4	42.9	94.0	68.4
Link	Vicuna-33B	70.7	29.9	50.3
	ChatGPT	11.6	94.3	52.9
	GPT-4	35.4	93.2	64.3
Doc	Vicuna-33B	81.6	17.9	49.8
	ChatGPT	34.7	73.2	54.0
	GPT-4	**88.3**	40.8	64.6
CEG	ChatGPT	40.1	79.2	59.7
	GPT-4	54.4	85.5	**69.9**

Table 2: Experimental results of our method powered by ChatGPT and GPT-4 on FELM worldknowledge subset. Baselines use the same GPT versions as our CEG, so the performances may vary from their original papers.

To summarize, CEG outperforms various SOTA baselines in two benchmarks with distinct LLM backbones, indicating that post-hoc retrieval with NLI is powerful for hallucination detection.

5.2 Results on Hallucination Regeneration

Method		Accuracy (%)
Baselines	Vanilla	63.40
	w/CoT	68.55
	w/Pre-Retrieval	61.35
CEG	w/ChatGPT	69.00
	w/GPT-3.5-Turbo-Instruct	**69.45**

Table 3: Experimental results powered by ChatGPT on the HaluEval QA subset. We employ two GPT models as the NLI method in our citation generation module.

On the HaluEval dataset, we adopt the full framework of CEG to further evaluate the regeneration module. If the doc is helpful in solving the problem and any of the response segments are classified into nonfactual, our method will generate a new prompt for regeneration as introduced in Section 3.4. Besides, a 2018 Wikipedia snapshot is adopted as the corpus (Gao et al., 2023a) in this experiment due to the inconsistency between this dataset and the previous corpus.

Experimental results are presented in Table 3. Firstly, our CEG framework with GPT-3.5-Turbo-Instruct achieves the best performance (69.45% in accuracy), which achieves 8.10% improvements compared to the pre-hoc retrieval strategy. CEG with ChatGPT also outperforms the pre-retrieval strategy, so these results demonstrate our post-hoc method is robust. Secondly, pre-hoc retrieval strategy even performs worse than the baseline with CoT (Li et al., 2023a), which indicates the retrieved documents are not always helpful. Thirdly, consistent with our experiments related to NLI models in Table 5, when using GPT-3.5-Turbo-Instruct as the NLI model, the results are superior to ChatGPT. We also conduct case studies to show our regeneration results, and some cases are shown in Appendix C.

5.3 Further Analyses

5.3.1 Ablation Study

Variants		Accuracy (%)		
		Nonfact	Factual	Balanced
ChatGPT	w/o RA	17.7	90.1	53.9
	w/o Threshold	40.1	78.2	59.2
	ALL	40.1	79.2	59.7
GPT-4	w/o RA	30.6	93.3	61.9
	w/o Threshold	50.3	83.6	67.0
	ALL	54.4	85.5	69.9

Table 4: Ablation results of CEG on the subset of FELM. 'RA' means the Retrieval Augmentation module.

We conduct ablation experiments on the FELM dataset, where the retrieval augmentation (k = 4) and the document selection threshold are involved (threshold = 0.5). As shown in Table 4, the retrieval augmentation module plays an important role in providing better results on different backbone LLMs (ChatGPT and GPT-4). Furthermore, the threshold of retrieved documents is necessary, which can filter out irrelevant documents in citation generation. Thus, all designed modules contribute to improvements in the CEG framework.

5.3.2 Retrieval Models

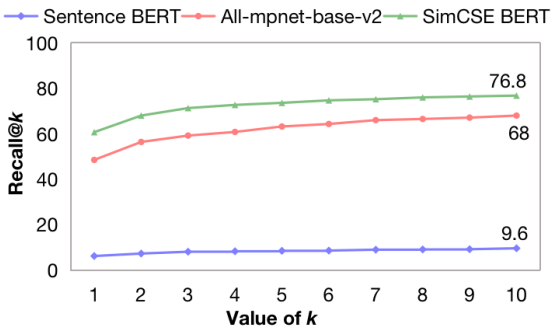

Figure 3: Performances of different retrievers on the WikiRetr-GPT3 dataset.

The choice of retrieval model significantly impacts the performance of our retrieval augmentation mod-

ule, as illustrated in Figure 3. Experimental results of three different retrieval models show that SimCSE BERT has better performances in post-hoc retrieval tasks (76.8% when using top 10 documents), where 64 million documents are used as candidates. Besides, although a larger value of k can improve recall, it also requires more resources for further computation. For a balance between efficiency and effectiveness, the value of k is set between 4 and 6 in our experiments.

5.3.3 NLI Models in Citation Generation

	WikiRetr-GPT3	WikiRetr-GPT4
True-9B	84	84
ChatGPT	66	77
GPT-3.5-Turbo-Instruct	**86**	91
GPT-4 Turbo	83	90
GPT-4	83	**96**

Table 5: Agreement rate (%) of different NLI models with human annotated instances on WikiRetr datasets.

The performance of different NLI models in the citation generation module is illustrated in Table 5, and there are two main observations we can make: 1) Despite being a state-of-the-art task-specific NLI approach, True-9B performs worse than best LLMs in this scenario. LLMs, such as GPT-3.5-Turbo-Instruct and GPT-4, are capable of playing the NLI model in our citation generation module, as they achieve high agreement rates with human annotators. 2) LLMs show better performance on data they generate, which is consistent with previous studies (Wang et al., 2023a; Zheng et al., 2023).

Metric	Precision (%)				Recall (%)	
NLI model	True		GPT-4		-	
Top-k	5	10	5	10	5	10
WikiRetr-GPT3	71.2	74.2	75.2	**78.2**	73.6	76.8
WikiRetr-GPT4	58.1	62.6	69.7	**75.7**	57.0	60.9

Table 6: NLI Precision of True-9B and GPT-4 on WikiRetr Datasets.

Table 6 shows experimental results of the citation generation module with distinct NLI models when the retriever is Simcse BERT, which indicate: 1) Even on the corpus with over 64 million candidates, our citation generation module exhibits outstanding performance, achieving 78.2 and 75.7 on the precision of the two datasets, respectively. 2) Compared to WikiRetr-GPT3, WikiRetr-GPT4 constitutes a more challenging and higher-quality dataset. The reason is that WikiRetr-GPT4 demonstrates lower recall, suggesting a lower semantic similarity between the original text and the generated claim. While its precision surpasses recall, indicating the generated claims are high quality.

5.3.4 Hyper-parameter Analysis

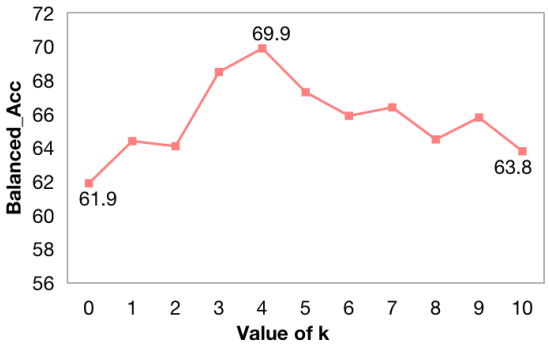

Figure 4: The variation of Balanced Accuracy with the change of k on the FELM dataset.

Due to the space limit, we only show the hyper-parameter analysis of k on FELM datasets in Figures 4. We have the following observations: 1) Experimental results on both FELM and WikiBio GPT-3 datasets demonstrate that more documents (larger value of k) do not always provide better performance. The reason can be more documents lead to longer input, attention becomes dispersed, and simultaneously, the relevance of the retrieved documents decreases. 2) Less or no retrieved documents also contribute to worse performance, which indicates the usefulness of the retrieval augmentation module in CEG. 3) The best performance achieved when k near 5 (Top-4 for FELM dataset and Top-6 for WikiBio GPT-3 dataset).

6 Conclusion

In this study, we propose a novel post-hoc citation-enhanced generation framework to reduce hallucinations in LLMs, which involves retrieval augmentation and natural language inference technologies. Different from previous hallucination studies, our framework is post-hoc and flexible, which can be applied to distinct LLMs without additional training or annotations, making it hold significant practical implications. Experiments on three hallucination-related benchmarks and our datasets demonstrate that our CEG framework achieves state-of-the-art performance in hallucination detection and regeneration. Further analyses show the effectiveness of our proposed modules and adopted models. In the future, we plan to further expand the corpus to support more scenarios.

Limitations

Our study has several limitations: 1) Restricted retriever and corpus: In our experiments, we do not employ a fine-tuned specific retriever for post-hoc methods, and our method utilizes only the Wikipedia corpus, limiting the applicability of our framework to general knowledge-based question-answering scenarios but only demonstrate the effectiveness of our model. 2) Our experiments are conducted on existing benchmarks and manual annotations, where we do not propose new QA datasets for the verification of regeneration performance. 3) The adopted NLI method in the citation generation module inherently relies on the LLM's world knowledge. More powerful NLI methods can contribute to better performance.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. *arXiv preprint arXiv:2310.00741*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Xiaohua Wang, Yuliang Yan, Longtao Huang, Xiaoqing Zheng, and Xuanjing Huang. 2023b. Hallucination detection for generative large language models by Bayesian sequential estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15361–15371, Singapore. Association for Computational Linguistics.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023c. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315, Singapore. Association for Computational Linguistics.

Yuxia Wang, Shimin Tao, Ning Xie, Hao Yang, Timothy Baldwin, and Karin Verspoor. 2023d. Collective human opinions in semantic textual similarity. *Transactions of the Association for Computational Linguistics*, 11:997–1013.

Wenhao Yu. 2022. Retrieval-augmented generation across heterogeneous knowledge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 52–58, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023a. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Appendix

A Prompts Used in Our Experiments

The prompt used in the evaluation of CEG and w/o threshold variants of CEG on the FELM dataset, as well as in the citation generation experiments on the HaluEval dataset, is presented in Table 11. The prompt for w/o retrieval augmentation variants of CEG in the FELM dataset evaluation is presented in Table 12.

The prompt to regenerate a new response on the HaluEval dataset is presented in Table 13.

- -

Context:
{ *Retrieved passages* }
Sentence:
{ *Sentence to be verified* }
Is the sentence supported by the context above? Answer Yes or No:

- -

The NLI prompt in experiments on WikiBio dataset is presented above.

- -

premise: { *Passage retrieved* } **hypothesis**: { *Hypothesis to be verified* }

- -

The NLI prompt employed by True-9B in experiments on WikiRetr datasets is presented above. We utilize the same prompt[6] as in Gao et al., 2023a. The agreement rates with human annotaters of True-9B in Table 5 and the precision of True-9B in Table 6 are based on this prompt.

- -

Premise: { *Passage retrieved* }
Hypothesis: { *Hypothesis to be verified* }
Task: Determine the logical relationship between premise and hypothesis.
Output format: If you believe the relationship is Entailment, output Entailment; for Contradiction, output Contradiction; for Neutral, output Neutral.

- -

The NLI prompt employed by GPT series models in experiments on WikiRetr datasets is presented above. The agreement rates with human annotaters of GPT models in Table 5 and the precision of GPT models in Table 6 are based on this prompt.

[6]You can find the original prompt at https://github.com/princeton-nlp/ALCE/blob/main/eval.py

B Detailed Information About Adopted Datasets

More Statistics of WikiBio GPT-3 dataset are shown in Table 7, and statistics of WorldKnowledge subset in FELM are summarized in Table 8.

#Passages	#Sentences	#Factual	#Nonfactual
238	1,908	516	1,392

Table 7: Statistics of WikiBio GPT-3 dataset.

	#Response	Error rate (%)	agreement rate (%)
Statistics	184	46.2	81.5
	#Segment	#Factual	#Nonfactual
Statistics	532	385	147

Table 8: Statistics of WorldKnowledge subset in FELM. The "Error rate" indicates the proportion of responses containing factual errors. The agreement rate is agreement rate between the two annotators during the annotation process.

In Table 9, we provide an example of HaluEval dataset.

#Knowledge#: The nine-mile byway starts south of Morehead, Kentucky and can be accessed by U.S. Highway 60. Morehead is a home rule-class city located along US 60 (the historic Midland Trail) and Interstate 64 in Rowan County, Kentucky, in the United States.
#Question#: What U.S Highway gives access to Zilpo Road, and is also known as Midland Trail?
#Right Answer#: U.S. Highway 60
#Hallucinated Answer#: U.S. Highway 70

Table 9: An example of HaluEval dataset.

The utilized modules in different datasets are summarized in Table 10.

Datasets	Retrieval	Citation	Regeneration
WikiBio GPT-3	✓	✓	-
FELM	✓	✓	-
HaluEval	✓	✓	✓
WikiRetr	✓	✓	-

Table 10: Modules of our framework used in four experimental datasets. "✓" indicates this module is adopted in the experiment, and "-" indicates not used.

C Case Studies on HaluEval Dataset

There are two cases in evaluation on HaluEval dataset, presented in Table 14 and Table 15, respectively.

Instruction: I will show you a question, a response segment of this question, and a reference doc. Your task is to assess whether the given response segment contains factual errors or not with the help of the reference doc. If you believe the segment contains factual errors, your answer should be "Nonfactual"; if there is no factual error in this segment, your answer should be "Factual". This means that the answer is "Nonfactual" only if there are some factual errors in the response segment. When there is no factual judgment in the response segment or the response segment has no clear meaning, your answer should be "Factual". Think it step by step. Give your reasoning first and then output the Answer.
Question:
{ *Question* }
Response segment:
{ *Response segment to be verified* }
Reference doc:
{ *Top-k docs concatenated with newline characters and numbers.* }
Output:

Table 11: Prompt used for the evaluation of CEG and w/o threshold variants of CEG on the FELM dataset, as well as in the NLI experiments on the HaluEval dataset.

Instruction: I will show you a question, a response segment of this question. Your task is to assess whether the given response segment contains factual errors or not. If you believe the segment contains factual errors, your answer should be "Nonfactual"; if there is no factual error in this segment, your answer should be "Factual". This means that the answer is "Nonfactual" only if there are some factual errors in the response segment.When there is no factual judgment in the response segment or the response segment has no clear meaning, your answer should be "Factual". Think it step by step. Give your reasoning first and then output the Answer.
Question:
{ *Question* }
Response segment:
{ *Response segment to be verified* }
Reference doc:
Output:

Table 12: Prompt used for the evaluation of w/o RA variant of our method on the FELM dataset.

User:
Instruction: I want you act as an answer judge. Given a question, two answers, your objective is to select the best and correct answer without hallucination and nonfactual information. You should try your best to select the best and correct answer. If the two answers are the same, you can randomly choose one. If both answers are incorrect, choose the better one. You MUST select an answer from the provided two answers. Think it step by step. Give your reasoning first and then output your choice. Output in the following format, "#Reasoning#:Your Reasoning\n#Choice#:Your Choice". Your choice should be "Answer 1" or "Answer 2".
#Question#: { *Question* }
#Answer 1#: { *Right_answer* }
#Answer 2#: { *Hallucinated_answer* }

Chatbot:
Reasoning and Answer

User:
Documents:
{ *Docs retrieved for the nonfactual segments* }
In your previous response, there are factual inaccuracies in the following passages:
{ *Nonfactual segments* }
Please re-answer the previous question with the help of documents. If the documents are unrelated to the issue, please ignore them. Output in the following format, "#Reasoning#:Your Reasoning\n#Choice#:Your Choice". Your choice should be "Answer 1" or "Answer 2".

Table 13: Prompt used for regeneration of our method on the HaluEval dataset.

#Question#: What film was written by the newspaperman born in 1880?
#Right_Answer#: The Lemon Drop Kid
#Hallucinated_Answer#: The Great American Novel written by Damon Runyon.

#Reasoning1#: Damon Runyon was a newspaperman born in 1880 and is known for writing The Great American Novel.
#Choice1#: Answer 2: The Great American Novel written by Damon Runyon.
#Segments#: 1.Damon Runyon was a newspaperman born in 1880 and is known for writing The Great American Novel.

#NLI Results#: [0,0]

#Doc Retrieved for Segment1#: Damon Runyon Alfred Damon Runyon (October 4, 1880 – December 10, 1946) was an American newspaperman and short-story writer. He was best known for his short stories celebrating the world of Broadway in New York City that grew out of the Prohibition era. To New Yorkers of his generation, a "Damon Runyon character" evoked a distinctive social type from the Brooklyn or Midtown demi-monde. The adjective "Runyonesque" refers to this type of character as well as to the type of situations and dialog that Runyon depicted. He spun humorous and sentimental tales of gamblers, hustlers, actors, and gangsters, few of

#Reasoning2#: Damon Runyon was known for his short stories celebrating the world of Broadway in New York City, not for writing The Great American Novel.
#Choice2#: Answer 1: The Lemon Drop Kid

Table 14: HaluEval Example 1. The red part is the content generated for the first time and annotated with citations. The green part is the content of the regeneration.

#Question#: Which low-cost airline is headquartered in Sydney?
#Right_Answer#: Jetstar Airways
#Hallucinated_Answer#: Tiger Airways is headquartered in Sydney.

#Reasoning1#: Jetstar Airways is a well-known low-cost airline and is headquartered in Melbourne, not Sydney. Tiger Airways used to be headquartered in Singapore, not Sydney.
#Choice1#: Answer 2: Tiger Airways is not headquartered in Sydney.
#Segments#: 1. Jetstar Airways is a well-known low-cost airline and is headquartered in Melbourne, not Sydney. 2. Tiger Airways used to be headquartered in Singapore, not Sydney

#NLI Results#: [0,0]

#Doc Retrieved for Segment1#: Jetstar Airways Jetstar Airways Pty Ltd, trading as Jetstar, is an Australian low-cost airline (self-described as \"value based\") headquartered in Melbourne. It is a wholly owned subsidiary of Qantas, created in response to the threat posed by low-cost airline Virgin Blue. Jetstar is part of Qantas' two brand strategy of having Qantas Airways for the premium full-service market and Jetstar for the low-cost market. Jetstar carries 8.5% of all passengers travelling in and out of Australia. The airline operates an extensive domestic network as well as regional and international services from its main base at Melbourne Airport, using a mixed

#Doc Retrieved for Segment2#: Australia Asia Airlines Australia Asia Airlines\"Àoyà Hángkōng Gōngsī\") was a wholly owned subsidiary of Qantas set up to operate services between Australia and Taiwan (Republic of China). The subsidiary was established due to the People's Republic of China objection to national carriers of countries with which it had diplomatic relations flying to a territory that it regarded as a breakaway province. The airline operated two Boeing 747SPs and a Boeing 767 aircraft seconded from the Qantas fleet, repainted in a modified livery, which did not display the Flag of Australia, or the kangaroo logo, which was replaced by

#Reasoning2#: According to the provided document, Jetstar Airways is headquartered in Melbourne, not Sydney. There is no mention of Tiger Airways being headquartered in Sydney.
#Choice2#: Answer 1: Jetstar Airways

Table 15: HaluEval Example 2. The red part is the content generated for the first time and annotated with citations. The green part is the content of the regeneration.