Controlling Risk of Retrieval-augmented Generation: A Counterfactual Prompting Framework

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) has emerged as a popular solution to mitigate the hallucination issues of large language models. However, existing studies on RAG seldom address the issue of predictive uncertainty, i.e., how likely it is that a RAG model's prediction is incorrect, resulting in uncontrollable risks in real-world applications. In this work, we emphasize the importance of risk control, ensuring that RAG models proactively refuse to answer questions with low confidence. Our research identifies two critical latent factors affecting RAG's confidence in its predictions: the quality of the retrieved results and the manner in which these results are utilized. To guide RAG models in assessing their own confidence based on these two latent factors, we develop a counterfactual prompting framework that induces the models to alter these factors and analyzes the effect on their answers. We also introduce a benchmarking procedure to collect answers with the option to abstain, facilitating a series of experiments. For evaluation, we introduce several risk-related metrics and the experimental results demonstrate the effectiveness of our approach. Our code and benchmark dataset are available at https://anonymous.4open.science/r/RC-RAG-0367.

1 Introduction

Large language models (LLMs) have gained considerable attention across a wide range of language tasks (Brown et al., 2020; Kandpal et al., 2023; Li et al., 2023b; Touvron et al., 2023). Despite the exciting performance, LLMs may suffer from hallucination issues (Ye et al., 2023; Azamfirei et al., 2023), due to limited memorization abilities or outdated pre-training corpora (Longpre et al., 2021; Xie et al., 2023). Recently, retrieval-augmented generation (RAG) has emerged as a promising solution to enhance factual accuracy (Kandpal et al., 2023; Xie et al., 2023; Gao et al., 2023), by synthesizing text snippets retrieved from external resources into final responses (Zhu et al., 2023; Ram et al., 2023; Izacard et al., 2023; Petroni et al., 2021; Ai et al., 2023).

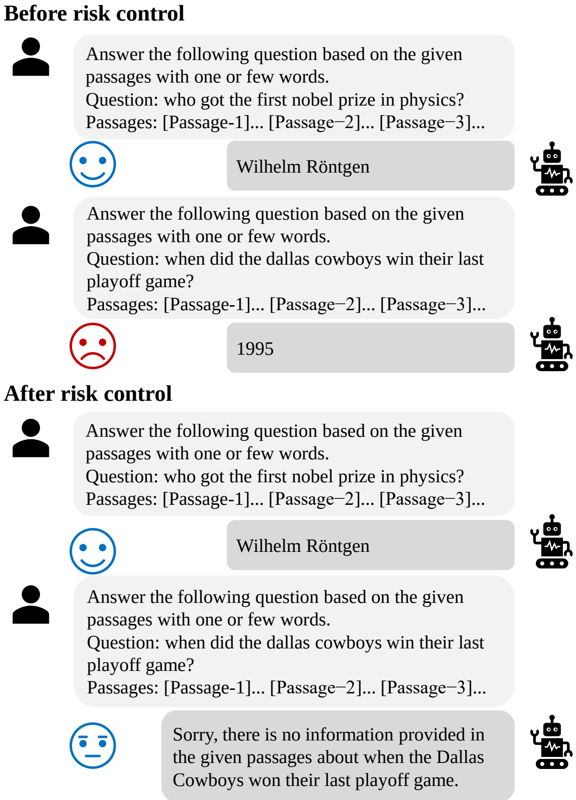

Figure 1: Illustration of risk control for RAG. Given a question, a risk controlled RAG model is expected to provide the correct answer if it has knowledge of the question, or alternatively, refuses to answer the question.

However, directly applying existing RAG techniques, particularly for knowledge-intensive tasks (Thorne et al., 2018; Yang et al., 2018; Petroni et al., 2021) such as factoid question answering (Aghaebrahimian and Jurcícek, 2016; Aghaebrahimian, 2018), introduces significant risks in practice. When confronted with noisy search results, even the most advanced RAG models are prone to producing unreliable answers, often exhibiting overconfidence in these erroneous responses (Yang et al., 2023; Ren et al., 2023). Such unreliable answers may severely undermine the user's question answering (QA) experience. Therefore, for practical applications, especially in sensitive domains like healthcare and legal assistance, it is crucial that RAG systems confidently provide answers when they know and state "I don't know" when they do not, as illustrated in Figure 1. This calls for the investigation on the risk control issue of RAG, a core research problem we want to tackle in this work. This approach reflects wisdom, as it involves RAG models proactively refusing to answer questions when predictions are uncertain.

Unfortunately, most previous research on risk control has focused on LLMs (Tian et al., 2023; Lin et al., 2023; Feng et al., 2024). There has been little work addressing the predictive uncertainty issue of RAG. Compared to the uncertainty assessment of LLMs, which concentrates on internal knowledge boundaries, the assessment for RAG requires additional consideration of external knowledge from retrieved results. In this work, we identify two critical factors during the uncertainty assessment of RAG: *the quality of the retrieved results and the manner in which they are used*. This raises an important research question: *how can we assess the predictive uncertainty of RAG based on these two retrieval results-related factors to determine when to discard or keep the generated answers*?

In this work, we propose a new task of risk control for RAG (RC-RAG) to decide whether to keep or discard the RAG outputs based on confidence assessment. We then introduce a novel *counterfactual prompting framework for RAG* under the zero-shot scenario, leveraging the counterfactual thinking for confidence assessment based on two latent factors. Counterfactual (Pearl, 2009) describes the human capacity to learn from prior experiences by imagining the outcomes of alternative actions that could have been taken. For a language model, we can inject counterfactual thinking into prompt, like "what if..." or "assume that", to imagine or simulate the consequences of changing a factor. Here, we induce the model to imagine scenarios where the quality of the retrieved results and their usage are poor, then measure its confidence based on the effect of these imagined scenarios on the answers. Specifically, our framework consists of three major modules, i.e., a prompting generation module, a judgment module, and a fusion module: (i) the prompting generation module generates answers under two scenarios that challenges the improper use and poor quality of the retrieved results, respectively; (ii) the judgment module determines whether to discard or keep the generated answers for both scenarios; and (iii) the fusion module combines the judgment results from both scenarios to produce the final decision for selective output. In order to avoid overestimating confidence affecting risk control, the prompting generation module and the judgment module are executed iteratively to produce the judgment for each scenario. It is important to note that our method is a general post-processing technique, making it applicable to almost any existing RAG method.

For evaluation, traditional metrics like Exact Match and F1 score typically focus on the effectiveness of RAG. In this work, we propose three risk-related metrics - *risk*, *carefulness*, *alignment*, and *coverage* - for risk-aware RAG evaluation. Due to the limited availability of datasets directly applicable to RC-RAG, we have constructed a novel risk control benchmark based on two publicly available QA datasets. Extensive experiments on RAG with Mistral (Jiang et al., 2024) and ChatGPT (Roumeliotis and Tselikas, 2023) as backbones demonstrate that the proposed framework can effectively abstain, outperforming baselines in 3 out of the 4 settings in terms of carefulness and risk, with up to a 14.76% improvement in carefulness and a 2.88% reduction in risk on average.

2 Related work

Retrieval-augmented generation. The typical retrieval-augmented generation (RAG) method follows a retrieve-then-generate pipeline, first retrieving relevant documents from a grounding corpus and then generating the final answer by the frozen generators (Shi et al., 2023; Ram et al., 2023). The retrieval augmentation is performed for all the questions through a single round (Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2021; Shi et al., 2023) or multiple rounds (Borgeaud et al., 2021; Ram et al., 2023; Trivedi et al., 2023; Jiang et al., 2023; Liu et al., 2024). However, such practice sometimes hurt generation performance, due to the unsatisfactory retrieved results (Mallen et al., 2023; Ren et al., 2023; Yoran et al., 2023; Tan et al., 2024). The reason may lie in the inconsistency between the relevance judgments in retrieval stage and the utility judgments in generation stage (Zhang et al., 2024). Besides jointly optimization of the retriever and generator (Guu et al., 2020; Lewis et al., 2020; Singh et al., 2021; Izacard et al., 2023), another solution is adaptive retrieval augmentation (Jiang et al., 2023; Asai et al., 2023; Wang et al., 2023), which actively determines when to retrieve based on self-knowledge.

Knowledge boundary. Detecting what LLMs know and do not know measures the boundary of models' internal knowledge, which can be applied to determine when to abstain it (Kadavath et al., 2022; Yang et al., 2023). The basic realization involves prompting one LLM to either verify in advance or to self-reflect on its response afterward (Ren et al., 2023; Li et al., 2024). It works for almost all LLMs, but there is a problem of overconfidence (Yin et al., 2023). Self-consistency between multiple inference also reflects the models' answering ability (Manakul et al., 2023), which is widely applicable but of high cost. Calibration-based methods obtain uncertainty or confidence scores of answers based on factors such as entropy, and token probability (Lin et al., 2023; Yang et al., 2023). A threshold is set to reject answers with low scores. Besides, some work elicits self-knowledge by referring to existing cases, which needs labeled samples. Through instruction tuning (Ouyang et al., 2022) or applying a small trainable model as classifier (Slobodkin et al., 2023; Azaria and Mitchell, 2023), LLMs can choose to abstain the answer when facing new questions. However, the limitation of the aforementioned work is that it only examines confidence when using internal knowledge, without considering the confidence when integrating external knowledge under the RAG setting. Though some work deals with knowledge conflict between the internal knowledge and external knowledge (Li et al., 2023a; Xie et al., 2023; Qian et al., 2023; Tan et al., 2024), it seldom rejects the RAG results, under the assumption that at least one kind of knowledge is true. This assumption is not conducive to risk control of RAG, since the retrieval results may contain noise. Therefore, in this work, we explore possible ways to control risk by discarding the RAG results, especially designed for external knowledge from retrieval results.

Counterfactual thinking. As the third level of the causal ladder after association and intervention, counterfactual reflects causality by imagining "what would the outcome be had the variable(s) been different" (Pearl, 2009; Nan et al., 2021). Counterfactual inference helps model unchanging causal mechanisms for better generalization and debias, which can be utilized for text classification, visual question answering, recommendation system and so on (Qian et al., 2021; Niu et al., 2021; Wei et al., 2021; Wang et al., 2022; Deng et al., 2023). It can calibrate causal effects through mediation analysis, by estimating the total effect and then eliminating the undesired effect (Xie et al., 2021). Different from these works, we focus on injecting counterfactual thinking into the prompt to better apply RaLLMs.

3 Problem statement

3.1 Task description

The RC-RAG task aims at assessing confidence or uncertainty of RAG answer to enable risk control in RAG. Formally, given a question Q and a group of retrieved passages P, the task outputs the answer A along with a judgment label $J \in \{0, 1\}$. For the samples with high confidence, the judgment label J is set as 1, indicating that the RAG answer could be *kept*. Oppositely, $J = 0$ is set for those uncertain output of RAG, which should be *discarded*.

Ideally, the assessment of confidence should align with the extent to which RAG knowledge supports the correct answer. In such cases, it can effectively mitigate risks by discarding answers that fall beyond the retrieval-augmented knowledge boundary, while still ensuring that correct answers are output as expected.

3.2 Benchmark

Data. To our best knowledge, there is limited available dataset that can be directly used for risk control for RAG. Therefore, we construct a RC-RAG benchmark composed of quadruple <Q, P, A, J> through simple automatic annotation. In the following, we introduce the data source and collection process of this benchmark.

Data source. In this work, we focus on factoid question answering (FQA) (Aghaebrahimian and Jurcícek, 2016; Aghaebrahimian, 2018), which typically provides a limited number of short answers, such as entities or numbers, and therefore carries a higher risk compared to non-factoid QA. We collect questions from two widely used datasets including Natural Questions (NQ) (Kwiatkowski et al., 2019) and TriviaQA (TQ) (Joshi et al., 2017). Since we focus on a zero-shot scenario, we collect question Q from their test sets.

Data collection. We further collect P, A, J based on questions Q in the data source.

	RC-TQ (7785)		**RC-NQ** (3610)	
	TQ-A	TQ-U	NQ-A	NQ-U
ChatGPT	5551	2234	1785	1825
Mistral	5553	2232	1830	1780

Table 1: Statistics of the full test sets and annotated results of answerable (A) and unanswerable (U) samples.

	Judgment result	
	Keep (K)	Discard(D)
Answerable (A)	AK	AD
Unanswerable (U)	UK	UD

Table 2: Categorization of the RAG output.

- **Passage collection.** For each question $q_i \in Q$, we utilize a dense retriever to retrieve top-k relevant passages $p_i = \{p_{i1}, ..., p_{ik}\}$ from external resources.

- **Answer generation.** Then, we prompt the LLM f to generate the answer $\hat{a}_i^f$ for each question-passage pair $\{q_i, p_i\}$, by feeding them as model input (prompts can be found in Appendix C.1):

$$\hat{a}_i^f = f(q_i, p_i). \tag{1}$$

- **Judgment annotation.** After that, we annotate j_i for each tuple of $\{q_i, p_i, \hat{a}_i^f\}$. As mentioned above, this judgment label indicates whether the RAG answer could be kept depending on confidence assessment. Ideally, samples with high confidence indeed belong to the answerable category, meaning their answers can be derived from the given knowledge. Since the evaluation of knowledge is non-trivial, we measure whether a sample is answerable approximately according to the correctness of the RAG answer $\hat{a}_i^f$, i.e.,

$$j_i = \begin{cases} 1, & \text{if } \hat{a}_i^f \text{ is correct,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

  The correctness can be measured based on the ground-truth answer a_i, through Exact Match (EM) score, F1 score and so on. Details can refer to Appendix A.

Finally, we obtain two RC-RAG datasets, i.e., RC-TQ and RC-NQ. The dataset statistics is shown in Table 1.

Evaluation. According to our RC-RAG benchmark, the samples could be divided into two cases, which are answerable (A) and unanswerable (U). Answerable ones refer to the samples whose RAG answer is correct, while unanswerable ones are the opposite. At the same time, there are two judgment results for RAG answers based on the designed judgment strategy, i.e., keep (K) and discard (D).

By combining above situations, the output of RAG would fall into one of the four folds, i.e., AK, AD, UD or UK, as shown in the Table 2. Specifically, AK/UK denotes the answerable/unanswerable samples with answers kept, while AD/UD denotes the answerable/unanswerable samples with answers discarded. Noted that samples answered wrongly are labeled as unanswerable ones based on our annotation, thus there is no case of keeping the wrong answer in the answerable samples.

Among these four folds, we further analyze which one causes the real risk in RAG. It is intuitive that the AK and UD folds pose no risk, as the judgment results are consistent with the labels. For AD fold, although the judgment result is inconsistent with the label, it poses no real risk since the user' behaviour may not be influenced when the RAG provides a null answer. Thus, only the UK fold exists risk, where the RAG sample is unanswerable but its answer is not discarded.

For evaluation, we propose four risk-aware evaluation metrics from various aspects, i.e., *risk*, *carefulness*, *alignment* and *coverage*.

- **Risk** (%) measures the percentage of risky cases (UK) among kept samples, i.e.,

$$risk = \frac{|\text{UK}|}{|\text{AK}| + |\text{UK}|},$$

  where || represents the number of samples.

- **Carefulness** (%) representing the percentage of incorrect samples being discarded, which is recall for unanswerable samples, i.e.,

$$carefulness = \frac{|\text{UD}|}{|\text{UK}| + |\text{UD}|}.$$

- **Alignment** (%) represents the percentage of samples where the judgment results are consistent with the labels, i.e.,

$$alignment = \frac{|\text{AK}| + |\text{UD}|}{|\text{AK}| + |\text{AD}| + |\text{UK}| + |\text{UD}|}.$$

- **Coverage** (%) measures the percentage of samples to be kept, i.e.,

$$coverage = \frac{|\text{AK}| + |\text{UK}|}{|\text{AK}| + |\text{AD}| + |\text{UK}| + |\text{UD}|}.$$

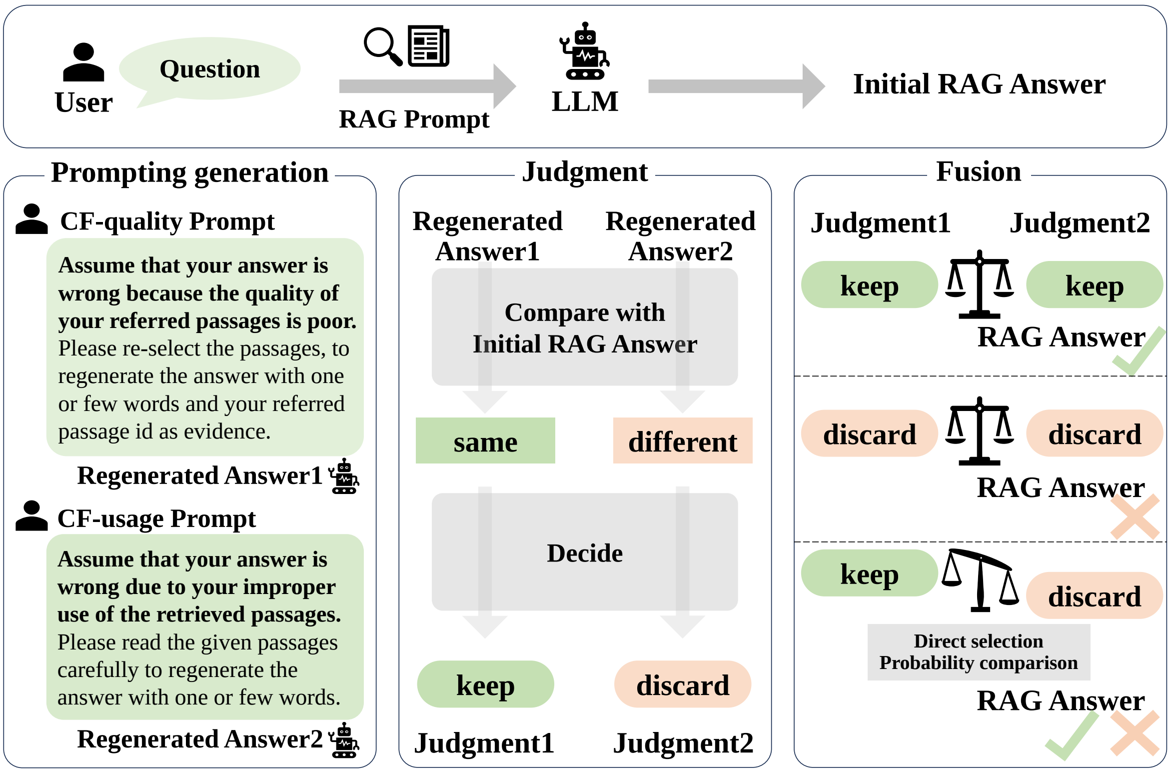

Figure 2: Overview of counterfactual prompting framework for RAG, in which the counterfactual (CF) prompts challenge the initial RAG answer in terms of the quality or usage of retrieved results. The final judgment result is derived from both aspects. Details refer to Sec. 4.

Note that a lower *risk* score is better, whereas higher scores are better for the other metrics.

4 Counterfactual prompting framework

Overview. To achieve risk control for RAG, we propose a novel counterfactual (CF) prompting framework that assesses predictive uncertainty of RAG. The overview is illustrated in Figure 2, consisting of a prompting generation module, a judgment module, and a fusion module: (i) a prompting generation module, which utilizes counterfactual thinking to induce answer regeneration effected by two changing factors; (ii) a judgment module, which makes judgment based on uncertainty assessment by analyzing the effect of each changing factor on their answer; and (iii) a fusion module for the final judgment result.

Prompting generation module. In this work, we assume that two latent factors can affect RAG uncertainty, i.e., the quality and the usage of retrieved results. Thus, we argue about each of them and ask for answer regeneration, respectively. Specifically, we implement each prompt as shown in Figure 2, where CF-quality prompt challenges the poor quality of retrieved results and CF-usage prompt challenges the improper usage. By imagining two scenarios that challenge each factor, the model adjusts the way it gets answers depending on its confidence level.

Judgment module. This module decides whether to keep or discard the answer according to uncertainty assessment for both scenarios. Specifically, we compare the regenerated answer with the initial RAG answer to analyze the effect of changing factors. There are two kinds of comparison results, i.e., same or different. Accordingly, the decision is made as follow: (i) *Keep*: Answer remaining the same indicates that the RAG answer is of relatively high confidence, which can be kept; (ii) *Discard*: Answer changing indicates that the RAG answer is uncertain, which should be discarded.

In order to avoid overestimating confidence, we iteratively execute the prompting generation module and the judgment module in N rounds to produce the judgment for each scenario. Specifically, if the answer remains the same after N rounds, it can be judged as *keep*. Otherwise, if the answer changes once, it is judged as *discard*.

Fusion module. We aggregate above judgment results as below. (i) If the two judgment results are consistent (both are *keep* or *discard*), we follow this judgment directly; (ii) Otherwise (one is *keep*, the other is *discard*), make the final judgment according to following prompts-based strategies (prompts can be found in Appendix C.3):

- **Direct selection:** We prompt the LLM to make a final decision, by telling it potential reasons resulting in wrong answers chosen from [*improper use* or *poor quality*] of retrieval results, according to the scenario in which the *discard* judgment was made in the previous judgment module.

- **Probability comparison:** We prompt the LLM to derive the probabilities of their respective judgments under two scenarios. By comparing the two probabilities, we select the judgment with the higher probability as the final judgment results.

After fusion, we change the judgment result of a special case from *keep* to *discard*: when the result is *keep* and the RAG output is "unknown". In this case, keeping the result of "unknown" is equivalent to discarding.

More details and the complete form of all the prompt can refer to Appendix B, C.

5 Experiment settings

Baselines. We compare our proposed CF prompting framework with three prompt-based baselines: (i) **If-or-Else (IoE) prompting framework** (Li et al., 2024), facilitating self-corrections based on LLMs' confidence. To adapt to the RC-RAG, we classify the case of answer correction as discard. (ii) **Calibration-based framework** (Tian et al., 2023), verbalizing confidence scores after obtaining answers, with a threshold set over verbalized scores. If the score is below the threshold, then choose to discard the output. (iii) **Priori judgement framework** (Ren et al., 2023), perceiving the factual knowledge boundary by self-judgment in the normal or RAG setting, which discards an answer by saying "unknown". More information about the baselines and their prompts can be found in Appendix D,E.

Backbones. We leverage two LLMs as backbones: Mistral (Jiang et al., 2024) and ChatGPT (Roumeliotis and Tselikas, 2023), which belong to open-source models and black-box models respectively. Note that these methods are general and can be extended to other LLMs.

Implementation details. For LLMs, we call OpenAI's API[1] to achieve ChatGPT (version gpt-3.5-turbo-0301), while we choose Mistral-7b[2] to implement Mistral. The max sequence length of LLM output is set to 256, and the temperature is set to 0. All the others are set as default. For the retrieved results, we conduct dense retrieval and sparse retrieval following Ren et al. (2023), and provide top-3 passages for each question following Wang et al. (2023). Most of the experimental results of our method use the direct selection fusion strategy, unless otherwise stated. More details refer to Appendix B. According to the analysis on the iteration number, as shown in Figure 3 in Appendix B, we report all results derived from a single run.

6 Experiment results

We aim to answer four research questions: **(RQ1)** Does our CF prompting framework efficiently control the risk of RAG compared with the baseline methods? **(RQ2)** Does the ability of LLMs affect the effectiveness of RC-RAG? **(RQ3)** Does the difficulty of QA task affect the ability of RC-RAG? **(RQ4)** Does the quality of retrieval results affect the effectiveness of RC-RAG? **(RQ5)** How does two CF-prompts affect the effectiveness of RC-RAG respectively? **(RQ6)** Are our risk control framework interpretable?

6.1 Main results

As shown in Table 3, we present the performance of different RC-RAG methods on two datasets. We have the following observations for **RQ1-3**.

Our approach effectively reduces risk and maintains carefulness compared to baselines. Baselines without a clear indication of the possible source of error struggle to reject uncertain RAG answers: (1) IOE has the worst rejection performance. For example, when using ChatGPT as a generator, it had the highest risk score and the lowest carefulness score on both datasets. This suggests that directly judging confidence in the answer is difficult to overcome the LLM's overconfidence problem in the RAG setting, due to reliance on retrieved results. (2) The calibration-based approach also suffers from overconfidence, resulting in the worst scores for risk and carefulness on both datasets when using Mistral as a generator. This shows that LLMs tend to output high confidence scores in the RAG setting without considering the potential misdirection of retrieved results. (3) The priori approach performs better on both metrics, particularly

[1] platform.openai.com

[2] huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Backbone	Method	RC-TQ				RC-NQ			
		risk↓	carefulness↑	alignment↑	coverage↑	risk↓	carefulness↑	alignment↑	coverage↑
Mistral	IoE	24.88	20.97	74.41	91.06	45.59	20.22	56.93	86.29
	Calibration	24.79	20.92	74.80	**91.47**	45.65	17.36	57.06	**89.25**
	Priori	21.95	33.87	**77.14**	86.38	42.61	28.60	61.52	82.63
	Ours	**19.00**	**52.87**	72.78	71.14	**38.22**	**52.98**	**63.60**	60.66
ChatGPT	IoE	21.59	33.53	78.88	**88.34**	41.79	31.14	64.29	**83.38**
	Calibration	19.71	42.51	79.45	83.75	40.97	35.34	64.96	79.78
	Priori	16.23	57.30	**79.68**	75.49	**34.72**	55.23	**70.55**	65.26
	Ours	**14.94**	**65.37**	75.38	66.55	35.22	**62.86**	66.23	53.24

Table 3: Main results of RC-RAG on the test set of two datasets and two LLMs with dense retriever. Best results in bold and second best in underline.

on the risk score of RC-NQ, achieving the lowest risk score of 34.72% with ChatGPT. This improvement is due to the prompt's mention of "based on the given information," leading the LLM to focus more on the quality of the retrieved results.

Our method outperforms the baselines in 3 out of the 4 settings (2 models and 2 datasets), achieving an average reduction of 2.88% on risk scores and an average improvement of 14.77% on carefulness scores. The results show that uncertainty prediction based on retrieval results explicitly can effectively help risk control. At the same time, alignment scores are not significantly inferior, especially on the RC-NQ dataset. However, as trade-off, the performance of coverage is inferior to the baseline method. It demonstrates how to balance risk control with coverage remains a difficult task.

Risk control ability is dependent on the LLM ability. We compare the performance of RC-RAG when using different LLMs as generators. We find that risk control works better with ChatGPT than with Mistral. Benchmark statistics (Table 1) show that Mistral outperforms ChatGPT on both datasets, particularly on RC-NQ. This indicates that risk control is more effective with weaker LLMs, underscoring the necessity of risk control methods. The underlying reason is that more capable models are more confident in both their internal knowledge and retrieved results. Consequently, Mistral achieves higher coverage scores, demonstrating that stronger LLMs tend to retain answers, which is consistent with the reasons for the above results.

Task difficulty has limited influence on risk control ability. We compared the effect of RAG risk control methods on different tasks. According to the risk and alignment scores, we find that the risk control methods perform worse in RC-NQ than in RC-TQ. The statistics of the benchmark (Table 1) show that RC-NQ is significantly more difficult than RC-TQ, as both ChatGPT and Mistral have a lower percentage of answerable samples on the RC-NQ dataset. We find that the more difficult the task to answer, the more difficult the risk control. For coverage scores, the performance in RC-NQ is also weaker. However, the performance in terms of carefulness scores was largely flat. The conclusion drawn from the above phenomenon is that *the difficulty of the task has a limited effect on the ability of the risk control method to accurately identify unanswerable samples*. As the proportion of samples that cannot be answered is larger in tasks with higher difficulty, the proportion of samples (UK) that cannot be answered but are retained will also be larger, and the risk and coverage scores will be correspondingly increased.

Method	risk↓	carefulness↑	alignment↑	coverage↑
Sparse retrieval				
IoE	65.18	30.55	47.73	75.18
Calibration	65.10	28.98	47.31	76.98
Priori	60.43	43.15	56.70	66.37
Ours	56.30	**65.80**	**65.15**	42.85
Dense retrieval				
IoE	45.59	20.22	56.93	86.29
Calibration	45.65	17.36	57.06	**89.25**
Priori	42.61	28.60	61.52	82.63
Ours	**38.22**	52.98	63.60	60.66

Table 4: Results of RC-RAG on the RC-NQ test set and Mistral with sparse retriever and dense retriever.

6.2 Impact of retriever

To answer **RQ4**, we compared the performance of risk control of RAG with different retriever. Results are shown in Table 4, conducted on the RC-NQ test set using Mistral as a generator.

By comparing the results using different retrievers, we observe that the risk control method is more cautious with the sparse retriever in terms of carefulness. However, due to the significantly lower number of unanswerable samples with the sparse retriever compared to the dense retriever, the risk score remains higher with the sparse retriever. Additionally, the experimental results show that our method outperforms all baselines using both retrievers in terms of risk, carefulness, and alignment.

6.3 Analysis of CF prompt and fusion strategy

To answer **RQ5**, we conduct ablation study to investigate the effects of the two CF prompts separately. The experiment was conducted on a subset of the RC-TQ test set using ChatGPT as a generator. First, we used CF-quality and CF-usage separately in prompting generation module, followed by the judgement module. Then for CF-quality, we analyze the results when the passage id is not required. The experimental results are shown in the Table 5, from which we have the following observations.

Only CF-usage prompting. The effect of risk control decreases while the Coverage score increases, indicating that the model tends to stick to its answer when confronted with challenge about the usage of retrieved results. This shows that the model is confident about the usage of retrieved results, which is essentially the internal knowledge of the LLMs, consistent with their characteristics of overconfidence.

Only CF-quality prompting. In contrast to the above, the risk score decreases significantly, indicating that the model tends to modify its answers when confronted with challenge about the quality of retrieved results. This shows that the model is sensitive to the challenge of the quality of retrieved results, which belongs to external knowledge, and the model itself does not have the ability to judge the quality of external knowledge.

No referred passage id provided. Although CF-quality was still used for prompting generation, the model did not update its answers if there was no need to provide a passage id. It indicates that the model may ignore the request to change the retrieved document and insist on its own answer when the passage id is not required.

Fusion strategy. The comparison results using two different fusion strategies are shown in Table 6 in Appendix F. Our complete approach with fusion module can effectively balance the two situations, considering both risk and coverage. Specifically, the direct fusion strategy can identify the unanswerable samples more effectively.

Method	risk↓	carefulness↑	alignment↑	coverage↑
Ours	13.56	75.75	76.00	59.00
only CF-usage	18.31	60.61	78.00	71.00
only CF-quality	10.48	83.33	74.50	52.50
w/o referred id	13.33	75.76	77.00	60.00

Table 5: Ablation study results of RC-RAG on the subset of RC-TQ test set and ChatGPT with dense retriever.

6.4 Case study

To answer **RQ6**, we conduct case study to demonstrate the working mechanism of our method.

As shown in Table 7 in Appendix G, we show the model prompting generation results and judgment results when facing different challenges, along with the final fusion results, which is based on ChatGPT augmented with dense retrieval. For the given question and retrieved result, the RAG answer and its referred retrieved passages do not correctly answer the question. However, none of baseline methods can correctly reject the answer. In our approach, the model still fails to recognize errors when the usage of retrieved results is challenged. When the quality of the retrieved results is challenged, the model realizes its reference does not effectively help answer the question, and ultimately chooses to reject the answer.

7 Conclusion

In this work, we propose a counterfactual prompting framework for assessing the uncertainty of RAG results, based on the quality of the retrieved results and the manner in which they are used. We construct a benchmark and design risk-related evaluation metrics. Experimental results with two LLMs on two datasets show that our method can effectively reject unanswerable samples and has a certain interpretability. In the future, we will explore other factors that may affect predictive uncertainty in RAG, such as conflicts between internal and external knowledge. Additionally, we will attempt to design objective functions based on risk-related metrics to guide the joint learning of the risk control framework and the RAG model.

Limitations

Our approach requires multiple rounds of prompting generation, which is computationally expensive. Therefore, further exploration of more efficient prompting generation methods is necessary. Additionally, the current fusion strategy is heuristic. In the future, semantic information can be incorporated to aid in the integration of the two judgments. Furthermore, we only considered risk control in a zero-shot scenario. In the future, designing objective functions based on risk-related metrics for joint training with the RAG framework could be explored, aiming to achieve an effective balance between risk control and answering capability.

Ethics statement

We have emphasized ethical considerations at every stage to ensure the responsible application of AI technologies. This work does not utilize personally identifiable information or require manually annotated datasets. Our methods are transparent, and we have made our data and code public to facilitate reproducibility and further research.

References

Ahmad Aghaebrahimian. 2018. Linguistically-based deep unstructured question answering. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 433–443.

Ahmad Aghaebrahimian and Filip Jurcícek. 2016. Open-domain factoid question answering via knowledge graph search. In *Proceedings of the workshop on human-computer question answering*, pages 22–28.

Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. 2023. Information retrieval meets large language models: a strategic report from chinese ir community. *AI Open*, 4:80–90.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):120.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2021. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xun Deng, Wenjie Wang, Fuli Feng, Hanwang Zhang, Xiangnan He, and Yong Liao. 2023. Counterfactual active learning for out-of-distribution generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11362–11377.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938.

Gautier Izacard and Édouard 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.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15696–15707.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Loka Li, Guangyi Chen, Yusheng Su, Zhenhao Chen, Yixuan Zhang, Eric Xing, and Kun Zhang. 2024. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *arXiv preprint arXiv:2402.12563*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, and Tianyu Du. 2024. Ra-isf: Learning to answer and understand from retrieval augmentation via iterative self-feedback. *arXiv preprint arXiv:2403.06840*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

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.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

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.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445.

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. 2023. " merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *arXiv preprint arXiv:2309.08594*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav

Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Konstantinos I Roumeliotis and Nikolaos D Tselikas. 2023. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *arXiv preprint arXiv:2401.11911*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

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*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. 2022. Causal representation learning for out-of-distribution recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 3562–3571.

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

Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1791–1800.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are large language models good at utility judgments? *arXiv preprint arXiv:2403.19216*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A Details about annotation

At the judgment annotation stage, we define the following criteria: Given the ground-truth answer a and the RAG answer $\hat{a}$, if $EM(a, \hat{a}) = 1, F1 > \tau, RougeL > \tau$, or the a appears in $\hat{a}$, the RAG answer can be judged as correct, and the sample can be annotated as answerable. We set $\tau = 0.7$.

B Implementation details

Details of judgment module. The criteria for determining that the answers remain unchanged are consistent with the criteria for matching the answers in the judgment annotation stage (Appendix A). If the regenerated answer matches the RAG answer, it can be judged as *same* and thus *keep*.

Details of iterative process. The number of our iterative process N is chosen from [1,2,3,4,5]. Specifically, we explored the performance of risk control when the number of iterations increased from 1 to 5, and the experimental setting was the same as Sec. 6.2. The results are shown in the figure 3, we can find that: with the increase of iterations, risk and coverage score showed a downward trend, carefulness score increased, while the alignment index was basically flat. In order to save the computational cost, we chose the number of iterations to be 1 to carry out the rest of our experiments.

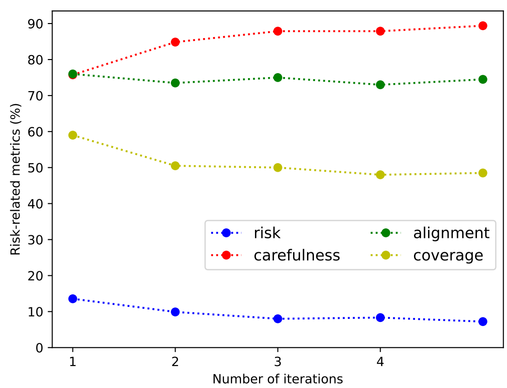

Figure 3: The change of risk-related metrics with the increase of iteration number.

C Prompt for CF prompting framework

C.1 Prompt for basic RAG setting

RAG prompt. *Answer the following question based on the given passages with one or few words. Provide your evidence between two ## symbols at the end of your response, either the passage id or your internal knowledge. For example, provide "Answer: apple. Evidence: ## Passage-0, Passage-1 ##." if you are referring to Passage-0 and Passage-1 to obtain the answer "apple". If there is no information in the passages, explain the answer by yourself.*

Question: {question}

Passages: {passage}

C.2 Prompt for prompting generation

CF-quality prompt. *Assume that your answer is wrong because the quality of your referred passages is poor. Please re-select the passages, to regenerate the answer with one or few words and your referred passage id as evidence.*

CF-usage prompt. *Assume that your answer is wrong due to your improper use of the retrieved passages. Please read the given passages carefully to regenerate the answer with one or few words.*

C.3 Prompt for fusion

Direct selection prompt.

- *Your answer is likely to be wrong because of the poor quality of retrieval passages, please choose to keep or discard this output. Generate $$ keep $$ if you choose to keep this answer, otherwise, generate $$ discard $$.*

- *Your answer is likely to be wrong because of the improper use of retrieval passages, please choose to keep or discard this output. Generate $$ keep $$ if you choose to keep this answer, otherwise, generate $$ discard $$.*

Probability comparison prompt. *Provide the probability that your regenerated answer is correct. Give ONLY the probability, no other words or explanation.*

For example:

Probability: <the probability between 0.0 and 1.0 that your specific answer is correct, without any extra commentary whatsoever; just the probability!>

Backbone	Method	RC-TQ				RC-NQ			
		risk↓	carefulness↑	alignment↑	coverage↑	risk↓	carefulness↑	alignment↑	coverage↑
Mistral	Ours$_{pro}$	21.23	43.37	72.68	**76.48**	41.72	44.49	60.17	**65.60**
	Ours$_{dir}$	**19.00**	**52.87**	**72.78**	71.14	**38.22**	**52.98**	**63.60**	60.66
ChatGPT	Ours$_{pro}$	16.30	59.96	**79.26**	**70.55**	36.24	57.87	**66.65**	**58.70**
	Ours$_{dir}$	**14.94**	**65.37**	75.38	66.55	**35.22**	**62.86**	66.23	53.24

Table 6: Comparison results of our methods using two different fusion strategies, on the test set of two datasets and two LLMs with dense retriever. The subscripts $_{dir}$ and $_{pro}$ represent the use of direct selection strategy and probability comparison strategy, respectively.

D Baselines

Among the three baseline methods, IoE and calibration-based framework are post-processing methods, while priori judgment framework is a pre-processing method.

IoE method was originally used for answer correction, requiring the model to update the answer of low confidence. If the model updates the answer, guide it to choose a final answer. Based on the matching results between the final answer and the RAG answer, we decide whether to keep or discard the RAG answer.

Calibration-based framework requires a threshold to discard answers. We set the threshold as 0.6 based on the experimental results.

Priori judgment framework requires prompt input only once, which explicitly mentions "given information" and "internal knowledge" in its prompt.

E Prompt for baselines

IOE prompt.

- *If you are very confident about your answer, maintain your answer. Otherwise, update your answer.*

- *You give two different answers in previous responses. Check the problem and your answers again, and give the best answer.*

Calibration prompt. *Provide the probability that your answer is correct. Give ONLY the probability, no other words or explanation.*

For example:

Probability: <the probability between 0.0 and 1.0 that your specific answer is correct, without any extra commentary whatsoever; just the probability!>

Priori prompt. *Given the following information:*

{passage}

Can you answer the following question based on the given information or your internal knowledge, if yes, you should give a short answer with one or few words, if no, you should answer "Unknown".

Question: {question}

F Analysis of fusion strategies

We show the comparison results of our methods using two different fusion strategies in Table 6.

G Case study

The generation result is shown in Table 7.

H AI Tool Usage Instructions

We utilized ChatGPT to assist in refining the expressions and wording of the paper.

Question: What was the purpose of a revival meeting during the second great awakening?
Ground-truth answer: Evangelizing.
Passage: [Passage-1]: Title: Second Great Awakening Content: a sense of personal salvation. Upon their return home, most converts joined or created small local churches, which grew rapidly. Calvinists emphasized the inability of men to save themselves. The only way to be saved was by God's electing grace. The Revival of 1800 in Logan County, Kentucky, began as a traditional Presbyterian sacramental occasion. The first informal camp meeting began there in June, when people began camping on the grounds of the Red River Meeting House. Subsequent meetings followed at the nearby Gasper River and Muddy River congregations, all three under the ministry of James McGready. One year later... [Passage-2]: ... [Passage-3]: ...

RAG answer: The purpose of a revival meeting during the Second Great Awakening was to inspire individuals to seek personal salvation and join or create small local churches. Evidence: Passage-1

IOE: The purpose of a revival meeting during the Second Great Awakening was to inspire individuals to seek personal salvation and join or create small local churches. Evidence: Passage-2, Passage-1 (Keep)
Calibration: Probability: 0.95 (Keep)
Priori: To inspire personal salvation and encourage individuals to turn to Christ.(Keep)

CF-usage: The purpose of a revival meeting during the Second Great Awakening was to inspire individuals to seek personal salvation and join or create small local churches, as well as to provide an opportunity for settlers to encounter organized religion and socialize with others. (Keep)
CF-quality: Sorry, there is no information provided in the given passages about the purpose of a revival meeting during the Second Great Awakening. (Discard)
Fusion: Discard

Table 7: An example (No.135) from the NQ test data, to analyze the generated answers and judgments of different risk control method for RAG. We mark the correct judgments in red and wrong ones in blue.