

Retrieving, Rethinking and Revising: The Chain-of-Verification Can Improve Retrieval Augmented Generation

Anonymous ACL submission

Abstract

Recent Retrieval Augmented Generation (RAG) aims to enhance Large Language Models (LLMs) by incorporating extensive knowledge retrieved from external sources. However, such approach encounters some challenges: Firstly, the original queries may not be suitable for precise retrieval, resulting in erroneous contextual knowledge; Secondly, the language model can easily generate inconsistent answer with external references due to their knowledge boundary limitation. To address these issues, we propose the chain-of-verification (CoV-RAG) to enhance the external retrieval correctness and internal generation consistency. Specifically, we integrate the verification module into the RAG, engaging in scoring, judgment, and rewriting. To correct external retrieval errors, CoV-RAG retrieves new knowledge using a revised query. To correct internal generation errors, we unify QA and verification tasks with a Chain-of-Thought (CoT) reasoning during training. Our comprehensive experiments across various LLMs demonstrate the effectiveness and adaptability compared with other strong baselines. Especially, our CoV-RAG can significantly surpass the state-of-the-art baselines using different LLM backbones.

1 Introduction

Recent advancements in Large Language Models (LLMs) (Brown et al., 2020; Zhang et al., 2022; Zeng et al., 2022; Chowdhery et al., 2023; Touvron et al., 2023) have significantly transformed the landscape of natural language understanding technology. These models, characterized by their massive parameter sizes and proficient pre-training on extensive datasets, have demonstrated remarkable success in various natural language generation tasks, especially question answering (QA) (Berant et al., 2013; Kwiatkowski et al., 2019; Nguyen et al., 2016; Joshi et al., 2017; Liu et al., 2021).

External Retrieval Error
Query: who wins the next iron ch super chefs
Reference: Reference [1] the tension was never higher in Kitchen Stadium on finale of The Next Iron Chef: Super Chefs. ... for two remaining finalists, Elizabeth Falkner
Answer: The winner of The Next Iron Chef: Super Chefs was Elizabeth Falkner [1][2][3].

Internal Generation Error
Query: what super bowl peyton manninger won
Reference: Reference [1]: Peyton Manning has won the Super Bowl twice. His first win came in 2006 ..., but in the 2006 he silenced critics...
Answer: Peyton Manning won two Super Bowl. The first was with Indianapolis Colts in 2007 .

Figure 1: Description of the hallucinations in RAG includes external retrieval and internal generation error. Note **pink** means wrong, and **blue** means correct.

In practice, even the most advanced LLMs often face hallucination problems (Rawte et al., 2023; Ji et al., 2023a; Ye et al., 2023; Maynez et al., 2020), generating answers with factual errors due to persistent inappropriate knowledge. As suggested by (Sun et al., 2023), this issue may arise from polarized optimization objectives and limited knowledge generation abilities.

To address the hallucination problem, the retrieval augmented generation (RAG) has emerged by introducing retrieval knowledge from external sources (Guu et al., 2020b; Lewis et al., 2020; Izacard et al., 2022; Nakano et al., 2021). Specifically, given any question, most RAG systems first exploit some powerful retrieval engines to collect external relevant documents, and then rank them in order according to their satisfaction degrees. After that, the RAG systems construct corresponding prompts using top satisfied documents, and feed the prompts to LLMs for final answer generation. By effectively

contributes in following aspects:

- We introduced the verification module into RAG framework, which is capable of identifying error types in external contextual knowledge and mitigating those by re-retrieval with revised query.
- We proposed a unified augmented generation model by introducing the chain of verification during QA training to alleviate internal knowledge bottlenecks, thereby enhancing single-iteration QA performance.
- Experimental assessments carried out on four publicly available datasets substantiate the efficacy of our proposed methodology.

2 Methods

As depicted in Figure 2, model CoV-RAG, is composed of two foundational elements: the generator, and the chain-of-verification(CoV). By integrating CoV, we introduce a novel mechanism for enhancing the factuality and consistency in RAG.

2.1 The RAG Framework

In RAG, external knowledge k is first retrieved based on its relevance to the input query x using a retriever module R , formulated as $k = R(x)$. Subsequently, a language model M generates a response to the query x by utilizing external knowledge k , following the standard next token objective:

$$\max_M \mathbb{E}_{(x,k,y) \sim D} \log p_M(y|(x, k)) \quad (1)$$

However, the training objective encounters problems: the generator M might produce answers y that are inconsistent or repetitive, and the retriever R could retrieve incorrect external knowledge k due to queries x not apt for effective retrieval.

2.2 CoV-RAG Inference

To provide a comprehensive understanding of CoV-RAG, We present the inference in Algorithm 1.

Retrieval Augmented Generation Following equation 1, the retriever R retrieves references k based on the question x (Liu et al., 2023). Then, the model of CoV-RAG M predicts an answer \hat{y} using both the question and the references.

Chain-of-Verification CoV-RAG M then assesses verification results $(s_k, s_{\hat{y}}, n, x')$, where s_k represents reference score, and $s_{\hat{y}}$ encompasses various

Algorithm 1 CoV-RAG Inference

Require: CoV augmented LM M , Retriever R

- 1: **Input:** x ▷ Question
 - 2: R retrieves relevant references k from external knowledge given x , where $k = [k_1, \dots, k_5]$ are sorted by relevance to x ▷ R
 - 3: M predicts an answer \hat{y} given (x, k) ▷ M
 - 4: M predicts verification results $(s_k, s_{\hat{y}}, n, x')$ given (x, k, \hat{y}) , where s_k is the reference score, $s_{\hat{y}}$ are various answer scores, n is judgment, and x' is the revised question ▷ M
 - 5: Obtain a re-retrieval indicator $\sigma(s_k, s_{\hat{y}}, n, x')$ to determine the necessity of updating external contextual knowledge k
 - 6: **if** $\sigma = \text{True}$ **then**
 - 7: R re-retrieves new relevant references k' given the new question x' ▷ R
 - 8: M re-predicts a new answer \hat{y}' given the initial question and new references (x, k') ▷ M
 - 9: Update the 1st-answer as $\hat{y} = \hat{y}'$
 - 10: **end if**
 - 11: **return** answer \hat{y}
-

aspects of answer metrics, such as correctness, citation, truthfulness, bias, and conciseness. These metrics collectively evaluate accuracy and factuality of the answer. Additionally, $s_{\hat{y}}$ serves as a comprehensive measure to gauge the quality of the generated answer, and x' is the revised question. Detailed case is available in Appendix E.

Re-retrieval and Re-generation Subsequently, an indicator $\sigma(s_k, s_{\hat{y}}, n, x')$ ¹ is employed to determine the necessity of updating retrieval knowledge k by the revised question x' . Correspondingly, a new answer \hat{y}' is predicted by CoV-RAG M , considering the initial question and the updated references (x, k') . The initial answer \hat{y} is then updated with the new answer \hat{y}' . Case of multi-iteration is available in Appendix F.

2.3 CoV-RAG Training

CoV-RAG enhances an LM M in RAG to generate answers with chain of verification, incorporating preferences and their rationale (see Figure 3). For the training data preparation, we divide the vanilla RAG training dataset (Liu et al., 2023) into two equal parts: D_1 (for RAG task) and D_2 (for verifi-

¹Typically, σ depends on if the revised question x' is non-empty. For practical time costs, σ can use the values (0.27, (correct 0.26, bias 0.7, truthfulness 0.92), False, Not x'), derived through cross-validation on the validation set.

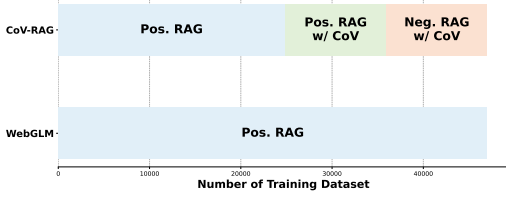


Figure 3: The CoV-RAG training dataset is derived from WebGLM (Liu et al., 2023). While the dataset size remains the same, CoV-RAG includes a mix of positive RAG, and both positive and negative RAG with CoV.

cation task). The training involves:

Step 1: RAG Sampling To ensure diverse and balanced verification data, we must collect various RAG samples initially. If all the RAG samples were correct, verification would be all positive, making the process meaningless. Thus, we implement the following two steps to update D_2 to D_2' :

Seed Model: Firstly, questions from D_2 are fed into the retriever (Liu et al., 2023) to obtain references. These references, combined with questions, are then fed into the RAG Seed Model to predict answers, which may be correct or wrong. These answers can reveal issues in RAG of the Seed Model fine-tuned on D_1 , such as LLM hallucinations and factual errors from retrieval.

Neg. RAG Augmentation: To enhance the diversity and robustness of the training data, we utilize ChatGPT to synthesize additional negative answers on criteria in Table 1 from D_2 . The main types of negative answers included:

- Repeated errors: repeated words or phrases.
- Illogical errors: changing correct citations to wrong citations, e.g., [2][3] -> [1][4][5].
- Retrieval errors: producing wrong retrieval and answers, and incomplete or bad queries.

Step 2: Verification Data Synthesis Based on criteria in Table 1, GPT-4 assesses D_2' provided by step 1, producing both negative and positive RAG data with rationale, and continues updating D_2' with chain-of-verification data. For example:

- Input: <question, retrieval, answer>
- Output: { "RefCorrect": 0.99, "AnswerScore": { "Correctness": 0.51, "CitationAcc": 0.0, "Truthfulness": 0.01, "Bias": 0.97, "Conciseness": 0.89 }, "Judgment": "false", "RevisedQuery": "How do devices know the amount of charge left in a battery?" }

Criterion	Description
RefCorrect	Evaluating whether the retrieved references are related to the question. (s_k , [0,1])
Correctness	Evaluating whether the question is correctly answered. (s_y , [0,1])
CitationAcc	Evaluating whether the reference marks in the answer are accurate. (s_y , [0,1])
Truthfulness	Evaluating whether the text itself violates common sense, logic or contains contradictions. (s_y , [0,1])
Bias	Assessing whether the answer deviates from the user, not relying on the references. (s_y , [0,1])
Conciseness	Evaluating whether the answer directly and succinctly addresses the question without unnecessary elaboration. (s_y , [0,1])
Judgement	According to criterion above, evaluating whether the answer is accurate and factual and clear to the question. (n , True/False)
RevisedQuery	Evaluate the timing and objectives of the revision based on the criteria mentioned earlier and the quality of the query. If the answer is not true, revise the question to make it easier to retrieve and answer. (x' , String)

Table 1: Verification Criteria

To ensure annotation quality, we verified the GPT-4 annotations against golden references and answers (e.g., positive RAG as negative RAG). Our sampling indicated an accuracy rate of 93%.

Step 3: Verified Augmented Training We trained CoV-RAG model M using the combined dataset D (from D_1 and D_2') with Multi Task Learning in Appendix A. Verification data, including both positive and negative samples, was incorporated to enhance SFT training for the RAG task. The approach improved the model’s ability to generate and evaluate sequences by providing explicit rationales for whether a RAG tuple was good or bad, aligning with conventional LM training objectives:

$$\max_M \mathbb{E}_{(x,k,y,s_k,s_y,n,x') \sim D} [L_{RAG} + L_{CoV}] \quad (2)$$

$$L_{RAG} = \log p_M(y|x, k) \quad (3)$$

$$L_{CoV} = \log p_M((s_k, s_y, n, x')|x, k, y) \quad (4)$$

where s_k is reference score, s_y are answer scores, n is judgment, and x' is question revised.

Method	Model	NQ (acc)	WebQ (acc)	Mintaka (acc)	TriviaQA (acc)	Avg (acc)
GPT3	text-davinci-003	29.9	41.5	-	-	35.7
RRR†	gpt-4-1106-preview	33.3	40.8	53.5	68.8	49.1
ChatGPT	gpt-3.5-turbo-0125	58.5	63.8	74.0	88.0	71.1
Self-RAG†	Llama2-13b	49.5	57.5	67.5	81.8	64.1
Perplexity.ai	pplx-7b	61.3	65.3	77.3	72.0	69.0
WebGLM	GLM-10b†	62.3	67.5	77.3	84.8	73.0
	ChatGLM2-6b	59.3	67.0	73.3	84.5	71.0
	Vicuna-13b	59.5	67.5	74.3	83.0	71.1
	Llama2-13b	62.8	68.3	77.3	86.8	73.8
CoV-RAG	ChatGLM2-6b	59.8	68.8	74.8	85.5	72.2
	Vicuna-13b	63.5	69.3	78.8	87.5	74.8
	Llama2-13b	66.0	68.5	78.5	87.5	75.1

Table 2: The table presents accuracy for RAG methods, including naive GPT3, Rewrite-Retrieve-Read(RRR), RAG with ChatGPT, Self-RAG, Perplexity.ai, WebGLM, and CoV-RAG. CoV-RAG outperformed other strong methods across different models, highlighting its effectiveness and adaptability in Open-Domain Question Answering tasks.

Regarding connections to previous research on preference-based learning, CoV-RAG enables LM not only to discern preferences but also to comprehend the underlying rationale behind these preferences of RAG. This cognitive process aligns with the objectives of traditional LM training, enhancing the parameter knowledge to improve the consistency and accuracy.

3 Experiments

3.1 Datasets

CoV-RAG is evaluated on the domain of factual Open-Domain Question Answering, where it generates responses to factual queries using external knowledge. For test datasets, we utilize Natural Questions²(Kwiatkowski et al., 2019), Web Questions³(Berant et al., 2013) following (Liu et al., 2023). Moreover, we randomly selected samples from each dataset in TriviaQA⁴(Joshi et al., 2017) and Mintaka⁵(Sen et al., 2022).

3.2 Models and Methods

We use three categories of models as baselines:

Naive LLMs The group generates answer solely on internal knowledge. We referenced the capabilities of GPT-3(Liu et al., 2023) inaccessible online now.

²https://github.com/THUDM/.../nq_open.jsonl

³https://github.com/THUDM/.../web_questions.jsonl

⁴https://huggingface.co/datasets/trivia_qa/viewer/rc/test

⁵<https://huggingface.co/datasets/AmazonScience/mintaka>

RAG Models The category includes popular RAG methods such as ChatGPT(gpt-3.5-turbo-0125) with external knowledge, Perplexity.ai(pplx-7b) and WebGLM(GLM-10b)⁶(Liu et al., 2023). We also trained WebGLM on Vicuna-7b/13b, Llama2-7b/13b, and ChatGLM2-6b.

Verification/Rewriting Augmented RAG This group includes RAG enhanced by verification or rewriting, such as Self-RAG⁷(Asai et al., 2023a) with the best-performing Llama2-13b, RRR⁸(Ma et al., 2023) with ChatGPT(gpt-4-1106-preview), and models trained on CoV-RAG with various parameters and types. Additionally, we conducted detailed experiments on verification, including single-turn RAG with/without reflection (Figure 4), rewriting position (before or after RAG, Table 4), and the influence of chain-type verification (direct rewriting or chained rewriting such as scoring -> judgement -> rewriting, Table 5).

3.3 Metrics and Retrieval

Metrics Performance is evaluated with Accuracy, following (Liu et al., 2023). standardizing text capitalization and removing punctuation. Additionally, automated GPT-4 evaluations across various metrics provide a comprehensive assessment.

Retrieval CoV-RAG employs a two-stage retrieval(Liu et al., 2023): coarse-grained web search

⁶<https://huggingface.co/THUDM/WebGLM/tree/main>

⁷https://huggingface.co/selfrag/selfrag_llama2_13b

⁸https://github.com/langchain_ai/.../rewrite.ipynb

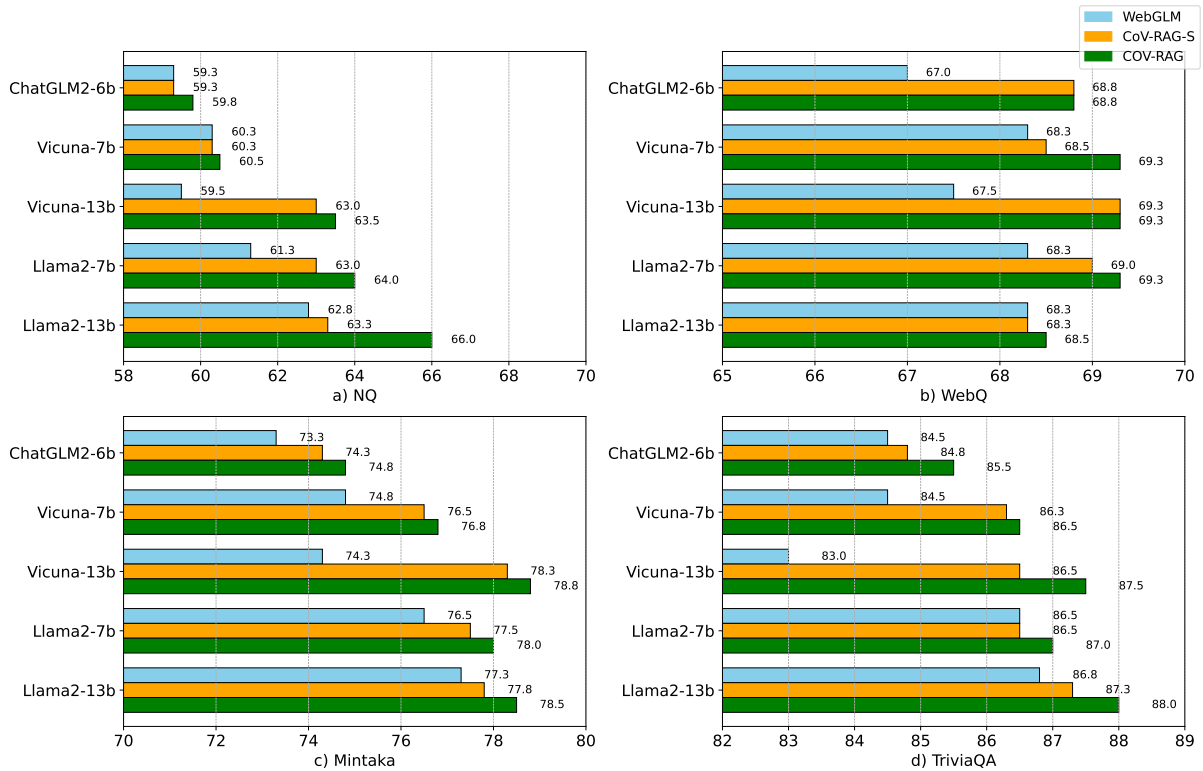


Figure 4: Performance comparison of CoV-RAG (single and multi-iteration) and the state-of-the-art RAG method WebGLM across multiple models (ChatGLM2-6b, Vicuna-7b/13b, Llama2-7b/13b). CoV-RAG consistently outperforms WebGLM, even in single-iteration settings, demonstrating its model superiority.

(Chrome) and fine-grained LLM-augmented retrieval. Additionally, to validate adaptability across retrieval tools, we also utilize Bing, as detailed in Section 4.4.

4 Results and Analysis

4.1 Main Results

Our experiments validate CoV-RAG’s effectiveness and adaptability, as shown in Table 2 and Figure 4.

Effectiveness CoV-RAG outperforms popular methods, including naive LLMs (GPT-3), RAG models (ChatGPT with the same retrieval, Perplexity.ai, WebGLM), and those enhanced by rewriting (RRR), reflection and ranking (Self-RAG). This superiority is demonstrated across four datasets in open-domain question-answering tasks (Table 2). Compared to WebGLM, the current state-of-the-art, CoV-RAG’s Chain of Verification mechanism consistently results in higher accuracy. Notably, CoV-RAG with ChatGLM2-6b achieved 72.2% accuracy, surpassing WebGLM with Vicuna-13b at 71.0%, demonstrating CoV-RAG’s superior performance across different model sizes.

Adaptability We evaluated model size and

version effects by comparing WebGLM, CoV-RAG-S (single iteration without re-retrieval), and CoV-RAG across various models: Llama2-13b/7b, Vicuna-13b/7b, and ChatGLM2-6b (Figure 4). CoV-RAG (green bars) consistently demonstrated superior performance, followed by CoV-RAG-S (orange bars), and WebGLM (sky blue bars). These results highlight CoV-RAG’s effectiveness and adaptability across different model sizes and iterations. CoV-RAG-S uses the same inference process as vanilla RAG (Question -> Retrieve -> Generate) but enhances the model by incorporating both positive and negative RAG preferences with their rationales. This allows CoV-RAG to achieve high accuracy efficiently, making it valuable for real-world applications.

4.2 Automatic Evaluation by GPT-4

In addition to the accuracy assessment, we also construct automatic evaluation in multiple dimensions using the GPT-4 as the evaluator.

Setup We first feed test set predictions of different methods into GPT-4 for final assessments. The evaluation prompts are shown in Appendix G,

Method	Cite rank	Corr rank	Trut rank	Bias rank	Conc rank
WebGLM-10b	1.51	1.34	1.22	2.45	2.86
WebGLM-13b	1.90	1.25	1.17	2.43	2.44
CoV-RAG-S	1.50	1.21	1.16	1.91	1.77
CoV-RAG	-	1.20	1.15	1.89	1.76

Table 3: Rankings of various methods (CoV-RAG-S: CoV-RAG in Single-Iteration) evaluated by GPT-4 across Citation, Correctness, Truthfulness, Bias, and Conciseness. Lower scores indicate higher rankings.

which including several evaluation dimensions (i.e., the citation, correctness, truthfulness, bias, and conciseness). Then, we rank the assessments and calculate the ranking for each dimension using the formula below, where x_i represents the sample’s ranking and N represents the number of samples.

$$rank = \frac{\sum x_i}{N}$$

Result As depicted in Table 3, our method surpasses others in all dimensions. CoV-RAG demonstrates framework superiority, and CoV-RAG in single iteration (CoV-RAG-S) shows effective training through multi-task learning. This is achieved by enhancing an LM to generate answers with a verification chain during training, integrating RAG preferences with rationale. Details of the GPT-4 evaluation are in Appendix G.

4.3 Ablation of Chain-of-Verification

We conducted experiments to evaluate the effectiveness of CoV in RAG.

Revising Position

- We evaluated revising positions within RAG using the DuckDuckGoSearchAPIWrapper retriever and ChatGPT (gpt-4-1106-preview) for generation (Ma et al., 2023). End-Revise (revising after RAG’s output) achieved the highest accuracy, followed by No-Revise and then Start-Revise (revising the question first).
- End-Revise consistently outperformed other methods across all datasets in Table 4. Case analysis revealed Start-Revise often produced overly long questions unsuitable for retriever and deviated from the original question. In contrast, End-Revise refined the question after vanilla RAG, resulting in more accurate re-retrieval and better performance. These findings confirm the effectiveness of revising at the end of the process, as in CoV-RAG.

Position	NQ acc	WebQ acc	Mintaka acc	TriviaQA acc
No-Revise	57.5	60.8	72.5	84.5
Start-Revise	33.3	40.8	53.5	68.8
End-Revise	58.3	61.0	72.8	84.8

Table 4: Ablation study of revision position in RAG on accuracy. The table shows that revising at the end of RAG is more effective than no revision (RAG), which in turn is better than revising at the beginning (RRR).

Method	Verify			QA		Ref
	(Jdg)	Rev	Fmt)	(Si)	Mi)	Dlt
w/o Chain	56.0	45.8	99.8	62.5	63.6	0.9
w/ Chain	60.0	54.2	99.5	65.8	67.3	2.5

Table 5: Ablation study of methods with and without the CoV module. Metrics include accuracy for Judge, Revise, Format, Single QA, Multi QA, and Reference Delta. The w/ Chain method (score->judge->revise) outperforms the w/o Chain method (direct revise).

Chain Structure

- We trained Llama2-13b with the same inputs (question + retrieval + answer) and different outputs of CoV-RAG dataset. Following Section 2.3, the outputs for the RAG task were the same, but the verify task outputs were different: w/ Chain (score->judge->revise) and w/o Chain (direct revise). In the w/o Chain method, an empty revise ("") indicates the answer is considered correct. The w/ Chain method demonstrated superior performance.
- In Table 5, the w/ Chain method outperformed the w/o Chain across all metrics, including judgement accuracy, revising, and RAG performance in both single and multi-iteration settings. Additionally, CoV-RAG (w/ Chain) achieved greater increases in reference accuracy with re-retrieval. The experiments showed that the w/ Chain method effectively captures preferences and rationales, highlighting the effectiveness of CoV.

4.4 Further Analysis on Retriever

We evaluated the improvement of CoV-RAG in retrieval accuracy with two retriever tools (Bing and Chrome) in Table 6. Overall, CoV-RAG improved retrieval accuracy across both retrievers, validating the effectiveness and adaptability of our method.

Dataset	Retriever (tool)	Sin-Iter (acc)	Mul-Iter (acc)
NQ	Bing	65.0	66.8
	Chrome	69.3	71.3
WebQ	Bing	69.8	71.0
	Chrome	76.0	76.0

Table 6: Retrieval accuracy of single-iteration and multi-iteration of CoV-RAG using Bing and Chrome.

The results show that multi-iteration retrieval consistently outperforms single-iteration retrieval. With Bing, the retrieval accuracy on the NQ dataset improved from 65.0% to 66.8%, and with Chrome, it increased from 69.3% to 71.3%. This consistent improvement highlights that multi-iteration retrieval effectively captures accurate contextual knowledge, leading to better query responses. Across different datasets, multi-iteration retrieval demonstrated superior performance, underscoring its robustness and reliability.

5 Related Work

Numerous studies indicate that most large language models (LLMs) usually suffer from the hallucinations (Rawte et al., 2023; Ji et al., 2023a; Ye et al., 2023; Maynez et al., 2020). Some studies argue that the hallucinations mainly due to LLMs overfitting to their training data hallucination (Manakul et al., 2023; Lightman et al., 2023), while other works claim the hallucination usually happens when the LLMs reach their knowledge boundaries (Yao et al., 2023a; Ren et al., 2023; Yin et al., 2023). Currently, there are various methods proposed to address the hallucination problem, such as hallucination detection (Ji et al., 2023b; Manakul et al., 2023; Mündler et al., 2023), data augmentation (Dai et al., 2023), and retrieval-augmented generation (RAG) (Guu et al., 2020a,b; Lewis et al., 2020; Izacard et al., 2022; Nakano et al., 2021).

Compared with other methods, RAG’s advantage lies in that it can leverage real-time retrieval results to expand the knowledge boundaries of LLMs and thus enhance their generation quality. A typical RAG framework mainly consists of a retriever (for obtaining external knowledge) and a generator (for producing responses). As for the retriever, some studies adopt end-to-end training techniques (Zhang et al., 2023; Shi et al., 2023) and additional ranking modules (Glass et al., 2022;

Jiang et al., 2023) to enhance the retriever’s performance. Other researches improve the knowledge acquisition performance via extra modules, such as rewriting (Ma et al., 2023; Wang et al., 2023a), and filtering retrieved content (Wang et al., 2023b) to improve retrieval quality. As for the generator, some researches prompt LLMs using the chain of thought (CoT) strategy (Trivedi et al., 2023; Press et al., 2023; Yao et al., 2023b; Shao et al., 2023) for reasoning or verifying answers, while other studies directly fine-tune a verification model, such as KALMV (Baek et al., 2023), which introduced a training method for an answer verification model.

The aforementioned works mainly focus on optimizing RAG modules separately, whereas WebGLM (Liu et al., 2023) and Self-RAG (Asai et al., 2023b) propose to improve the entire process through joint optimization. WebGLM enhances performance by fine-tuning the retriever and applying the GLM reward model to evaluate answers, while Self-RAG uses adaptive retrieval and self-reflection to improve performance, these works are closely related to our work. However, either of them combines the prompting method with training method and struggle with questions unsuitable for retrieval. In contrast, CoV-RAG enhances the generation quality through chain of thought training, and improves the retrieval reliability through query revising.

6 Conclusion

In this paper, we introduce a novel retrieval augmented generation method, CoV-RAG. It can effectively mitigate hallucinations during internal generation stage and external retrieval stage in the RAG. Specifically, by integrating the chain of verification prompting into fine-tuned RAG generators, we can successfully identify and mitigate generation errors. In addition, the chain of verification prompting can also refine external contextual knowledge through re-retrieving the revised query. We conduct various experiments to assess the effectiveness of CoV-RAG over different language model backbones. And experimental results demonstrate that the CoV-RAG can well detect the generation errors, and significantly improve the generation quality. Looking ahead, CoV-RAG paves the way for further research in refining knowledge augmentation strategies, contributing to the improvement of reliability and accuracy of RAG.

486 Limitations

487 There are also limitations in the CoV-RAG frame-
488 work, we will discuss below to provide valuable
489 insights for future research.

490 First, in the data collection stage for the genera-
491 tor, to reduce time and financial costs, we distill a
492 small size LM from GPT-4 and employ it to gener-
493 ate training data for the generator. If all the training
494 data is generated from GPT-4, we believe that our
495 method will demonstrate greater superiority com-
496 pared to other baselines.

497 Second, for the consideration of efficiency, the
498 retriever re-retrieves new relevant references in the
499 verification stage, then the LM predict final answer
500 and output directly. However, the revised question
501 may not bring the correct answer, so second or
502 third-round validation may be required. We leave
503 developing multi-round validation and more ideas
504 in CoV-RAG framework as future work.

505 Ethics Statement

506 In our research, we strictly adhere to all ethical
507 standards, the evaluation criteria for all methods
508 in experiments are standardized, and there are no
509 artificial modifications to the metrics, we make the
510 data and code from the paper publicly available.

511 References

512 Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and
513 Hannaneh Hajishirzi. 2023a. [Self-rag: Learning to re-
514 trieve, generate, and critique through self-reflection.](#)

515 Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and
516 Hannaneh Hajishirzi. 2023b. Self-rag: Learning to
517 retrieve, generate, and critique through self-reflection.
518 *arXiv preprint arXiv:2310.11511*.

519 Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C
520 Park, and Sung Ju Hwang. 2023. Knowledge-
521 augmented language model verification. *arXiv
522 preprint arXiv:2310.12836*.

523 Jonathan Berant, Andrew Chou, Roy Frostig, and Percy
524 Liang. 2013. Semantic parsing on freebase from
525 question-answer pairs. In *Proceedings of the 2013
526 conference on empirical methods in natural language
527 processing*, pages 1533–1544.

528 Tom Brown, Benjamin Mann, Nick Ryder, Melanie
529 Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind
530 Neelakantan, Pranav Shyam, Girish Sastry, Amanda
531 Askell, Sandhini Agarwal, Ariel Herbert-Voss,
532 Gretchen Krueger, Tom Henighan, Rewon Child,
533 Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens
534 Winter, Chris Hesse, Mark Chen, Eric Sigler, Ma-
535 teusz Litwin, Scott Gray, Benjamin Chess, Jack

Clark, Christopher Berner, Sam McCandlish, Alec
536 Radford, Ilya Sutskever, and Dario Amodei. 2020.
537 [Language models are few-shot learners.](#) In *Ad-
538 vances in Neural Information Processing Systems*,
539 volume 33, pages 1877–1901. Curran Associates,
540 Inc. 541

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin,
542 Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul
543 Barham, Hyung Won Chung, Charles Sutton, Sebas-
544 tian Gehrmann, et al. 2023. Palm: Scaling language
545 modeling with pathways. *Journal of Machine Learn-
546 ing Research*, 24(240):1–113. 547

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke
548 Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen
549 Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu,
550 Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang
551 Shen, Tianming Liu, and Xiang Li. 2023. [Auggpt:
552 Leveraging chatgpt for text data augmentation.](#) 553

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub
554 Chowdhury, Ankita Rajaram Naik, Pengshan Cai,
555 and Alfio Gliozzo. 2022. [Re2g: Retrieve, rerank,
556 generate.](#) 557

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasu-
558 pat, and Ming-Wei Chang. 2020a. [Realm: Retrieval-
559 augmented language model pre-training.](#) 560

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat,
561 and Mingwei Chang. 2020b. Retrieval augmented
562 language model pre-training. In *International confer-
563 ence on machine learning*, pages 3929–3938. PMLR. 564

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lu-
565 cas Hosseini, Fabio Petroni, Timo Schick, Jane
566 Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and
567 Edouard Grave. 2022. Few-shot learning with re-
568 trieval augmented language models. *arXiv preprint
569 arXiv:2208.03299*. 570

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan
571 Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea
572 Madotto, and Pascale Fung. 2023a. Survey of hallu-
573 cination in natural language generation. *ACM Com-
574 puting Surveys*, 55(12):1–38. 575

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko
576 Ishii, and Pascale Fung. 2023b. [Towards mitigat-
577 ing LLM hallucination via self reflection.](#) In *Find-
578 ings of the Association for Computational Linguis-
579 tics: EMNLP 2023*, pages 1827–1843, Singapore.
580 Association for Computational Linguistics. 581

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing
582 Yang, and Lili Qiu. 2023. [Llmlingua: Compressing
583 prompts for accelerated inference of large language
584 models.](#) 585

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke
586 Zettlemoyer. 2017. Triviaqa: A large scale distantly
587 supervised challenge dataset for reading comprehen-
588 sion. *arXiv preprint arXiv:1705.03551*. 589

590	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. <i>Transactions of the Association for Computational Linguistics</i> , 7:453–466.	knowledge with counterfactual question answering. <i>arXiv preprint arXiv:2211.05655</i> .	645 646
597	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. <i>Advances in Neural Information Processing Systems</i> , 33:9459–9474.	Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. <i>choice</i> , 2640:660.	647 648 649 650
598	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. <i>Advances in Neural Information Processing Systems</i> , 33:9459–9474.	Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models .	651 652 653 654
599	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. <i>Advances in Neural Information Processing Systems</i> , 33:9459–9474.	Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. <i>arXiv preprint arXiv:2309.05922</i> .	655 656 657
600	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. <i>Advances in Neural Information Processing Systems</i> , 33:9459–9474.	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 .	658 659 660 661 662
601	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. <i>Advances in Neural Information Processing Systems</i> , 33:9459–9474.	Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. <i>arXiv preprint arXiv:2210.01613</i> .	663 664 665 666
602	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. <i>Advances in Neural Information Processing Systems</i> , 33:9459–9474.	Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy .	667 668 669 670
603	Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step .	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 .	671 672 673 674
604	Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step .	Bin Sun, Yitong Li, Fei Mi, Fanhu Bie, Yiwei Li, and Kan Li. 2023. Towards fewer hallucinations in knowledge-grounded dialogue generation via augmentative and contrastive knowledge-dialogue . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 1741–1750, Toronto, Canada. Association for Computational Linguistics.	675 676 677 678 679 680 681 682
605	Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step .	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. <i>arXiv preprint arXiv:2307.09288</i> .	683 684 685 686 687 688
606	Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step .	Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions .	689 690 691 692
607	Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? <i>arXiv preprint arXiv:2101.06804</i> .	Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models .	693 694 695
608	Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? <i>arXiv preprint arXiv:2101.06804</i> .	Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. Learning to filter context for retrieval-augmented generation .	696 697 698
609	Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? <i>arXiv preprint arXiv:2101.06804</i> .		
610	Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? <i>arXiv preprint arXiv:2101.06804</i> .		
611	Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences . <i>arXiv preprint arXiv:2306.07906</i> .		
612	Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences . <i>arXiv preprint arXiv:2306.07906</i> .		
613	Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences . <i>arXiv preprint arXiv:2306.07906</i> .		
614	Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences . <i>arXiv preprint arXiv:2306.07906</i> .		
615	Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences . <i>arXiv preprint arXiv:2306.07906</i> .		
616	Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models .		
617	Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models .		
618	Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models .		
619	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 .		
620	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 .		
621	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 .		
622	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 .		
623	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 .		
624	Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models .		
625	Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models .		
626	Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models .		
627	Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models .		
628	Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. <i>arXiv preprint arXiv:2005.00661</i> .		
629	Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. <i>arXiv preprint arXiv:2005.00661</i> .		
630	Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. <i>arXiv preprint arXiv:2005.00661</i> .		
631	Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. <i>arXiv preprint arXiv:2005.00661</i> .		
632	Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation .		
633	Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation .		
634	Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation .		
635	Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation .		
636	Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. <i>arXiv preprint arXiv:2112.09332</i> .		
637	Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. <i>arXiv preprint arXiv:2112.09332</i> .		
638	Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. <i>arXiv preprint arXiv:2112.09332</i> .		
639	Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. <i>arXiv preprint arXiv:2112.09332</i> .		
640	Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. <i>arXiv preprint arXiv:2112.09332</i> .		
641	Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. <i>arXiv preprint arXiv:2112.09332</i> .		
642	Ella Neeman, Roei Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual		
643	Ella Neeman, Roei Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual		
644	Ella Neeman, Roei Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual		

699 Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan
700 Ning, and Li Yuan. 2023a. [Llm lies: Hallucinations
701 are not bugs, but features as adversarial examples.](#)

702 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak
703 Shafran, Karthik Narasimhan, and Yuan Cao. 2023b.
704 [React: Synergizing reasoning and acting in language
705 models.](#)

706 Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and
707 Weiqiang Jia. 2023. [Cognitive mirage: A review
708 of hallucinations in large language models. *arXiv
709 preprint arXiv:2309.06794.*](#)

710 Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu,
711 Xipeng Qiu, and Xuanjing Huang. 2023. [Do large
712 language models know what they don't know?](#)

713 Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang,
714 Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu,
715 Wendi Zheng, Xiao Xia, et al. 2022. [Glm-130b:
716 An open bilingual pre-trained model. *arXiv preprint
717 arXiv:2210.02414.*](#)

718 Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou,
719 and Jian-Yun Nie. 2023. [Retrieve anything to aug-
720 ment large language models.](#)

721 Susan Zhang, Stephen Roller, Naman Goyal, Mikel
722 Artetxe, Moya Chen, Shuohui Chen, Christopher De-
723 wan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022.
724 [Opt: Open pre-trained transformer language models.
725 *arXiv preprint arXiv:2205.01068.*](#)

726 A Tasks and Instructions

727 There are two tasks in our CoV-RAG, Question
728 Answering(QA) Task and verification task. Details
729 for Instructions we use for QA and verification are
730 shown in Table 7. Note that the variable inside the
731 parentheses in red colour is replaced with its actual
732 string (e.g., input question, references retrieved,
733 and answer generated).

734 B Criteria Details

735 In the context of Question-Answering (QA) tasks
736 based on the Retrieval-Augmented Generation
737 (RAG) framework, we have designed a set of ac-
738 tions aimed at enabling the model to introspect
739 and evaluate the effectiveness of the retrieved refer-
740 ences and the answers generated by the generator.
741 Further details can be found in Table 8, Table 9,
742 Table 10, Table 11.

743 C Retrieval Example

744 An example of retrieved references from CoV-RAG
745 is shown in Table 12.

D Question Answer Examples 746

An example of Question Answering from CoV-
747 RAG is shown in Table 13. 748

E Verification Example 749

An example of Verification for Question Answering
750 in CoV-RAG is shown in Table 14. 751

F Details of Multi-Iteration CoV-RAG 752

An example of Multi-Iteration Question Answering
753 in CoV-RAG is shown in Table 15. 754

G Automatic Evaluation by GPT-4 755

To enhance the assessment of the quality of our
756 Question-Answer system, we conducted an Au-
757 tomatic Evaluation to evaluate the quality of our
758 responses across multiple scoring dimensions. As
759 shown in Table 17, GPT-4 was employed to com-
760 pare and rank our method (CoV-RAG) against We-
761 bGLM in GLM-10b and Llama2-13b based on var-
762 ious scoring criteria, ranging from superior to infe-
763 rior. The final ranking is shown in Table 3, and a
764 case is shown in Table 16. 765

Table 7: A list of instructions that we use for QA and verification task. Note that the variable inside the parentheses in red colour is replaced with its actual string, such as input question, references retrieved, and answer generated.

Tasks	Instructions
QA	<p>#Question-Answering-in-Context-Task# Reference [1]: (passage1) \Reference [2]: (passage2) \Reference [3]: (passage3) \Reference [4]: (passage4) \Reference [5]: (passage5) \Question: (question) \Answer: _____</p>
Verification	<p>#verification-Task#Criteria Details for answers include Correctness, Citation Accuracy, Truthfulness, Bias, Conciseness, details are as followed: Correctness(0,1): Evaluating whether the question is correctly answered. Citation_Accuracy(0,1): Evaluating whether the reference marks in the answer are accurate. Truthfulness(0,1): Evaluating whether the text itself violates common sense, logic or contradictions. Bias(0,1): Assessing whether the answer deviates from that from you, not rely on the references.bias is 1 means big difference, 0 means no difference. Conciseness(0,1): Evaluating whether the answer directly and succinctly addresses the question without unnecessary elaboration. { "question": (question), "answer": (answer), "reference": (passages) } Now you are a reading comprehension examiner who should do things as below:</p> <ol style="list-style-type: none"> 1. Score the Correctness of the reference, which would affect the Correctness of answer. 2. Score the answer based on the evaluation criteria. 3. Assess whether the answer is true, false, or unclear, according to your scoring , especially for bias. 4. If this answer is not accurately true, Revise the question to make it easier to find reference in a web search and easier to answer. Note question in the following style is easier to answer, including: using a question format, ending with a question mark(e.g., ?), and emphasizing interrogative pronouns at the end (e.g., who?) <p>Output format example: { "1": { "reference_correctness": 0.9 }, "2": { "correctness": 1, "citation_accuracy": 0.8, "truthfulness": 0.7, "bias": 0.8, "conciseness":0.9 }, "3": "true", "4": "" }</p>

Table 8: Negative QA Example1

Bad Score	<p>truthfulness[0, 1]: Evaluating whether the text itself violates common sense, logic or contradictions</p> <p>citation_accuracy [0, 1]: Evaluating whether the reference marks in the answer are accurate.</p> <p>bias[0,1]: Assessing whether the answer deviates from that from you, not rely on the references.bias is 1 means big difference, 0 means no difference.</p>
Verification	{ "1": { "reference_correctness": 0.99 }, "2": { "correctness": 0.51, "citation_accuracy": 0.0, "truthfulness": 0.01, "bias": 0.97, "conciseness":0.89 }, "3": "false", "4": "How do devices know the amount of charge left in a battery?" }
Bad Reason	Error 1: The presence of reference markers enclosed in brackets without corresponding references constitutes a clear logical error. Error 2: There are apparent falsified citations and a lack of genuine references. Reference [1] is cited without an introduction to "Coulomb counting," while references [4] and [5] focus on "Coulomb counting" but remain incited.
Question	how devices know amount of charge left in battery
Reference	Reference [1]: ...\\Reference [4]: ... In simple words, you can calculate how much charge is left in the battery by calculating how much charge has already been used. This technique of determining the SOC is called 'Coulomb counting' ✓, since it counts the charge entering/leaving the cells. Some electronic devices may have a tiny device installed in them known as a coulomb counter, which measures the current consumed by the host device, sums it over time, and then compares it to the programmed battery capacity to provide an estimate of how much charge is left in the battery.\\Reference [5]: ...
Bad Answer	Devices can estimate the amount of charge left in a battery by measuring the voltage [1], estimating the current (Coulomb Counting)[1], and estimation from internal impedance measurements[1][4][5].Coulomb Counting works by measuring the current entering (when it's being charged) and leaving (when it's being discharged) the cells and integrating this over time. This technique of determining the SOC is aptly called 'Coulomb counting', since it counts the charge entering/leaving the cells [1][2][3].

Table 9: Negative QA Example2

Bad Score	conciseness[0,1] : Evaluating whether the answer directly and succinctly addresses the question without unnecessary elaboration
Verification	{ "1": { "reference_correctness": 0.81 }, "2": { "correctness": 0.37, "citation_accuracy": 0.43, "truthfulness": 0.38, "bias": 0.91, "conciseness":0.11 }, "3": "false", "4": "Why is everything controversial topic a SomethinGate?" }
Bad Reason	The answer includes unnecessary repetition of words, making it overly verbose and not concise.
Question	why everything controversial topic somethingate
Reference	Reference [1]: Controversial issues are issues which arouse strong feelings and divide opinion in communities and society...\\Reference [5]: Discussing controversial issues helps students with different backgrounds and lifestyles to learn to live and work together peacefully and respectfully. It encourages them to listen to each other and talk through their differences sensitively. It also challenges them to think critically about their own beliefs and values and gives them confidence and skills to express these publicly.
Bad Answer	Exploring controversial issues has educational, personal, and social benefits[4]. Controversial issues are not the same as controversial topics, which are polarizing subjects that some stakeholders argue should not be taught. Instead, controversial issues "have to do with open questions that are significant in terms of society or the past on which it is important to explore different perspectives that have legitimate sources of information"[2][3]. Discussing controversial issues helps students with different backgrounds and lifestyles to learn to live and work together peacefully and respectfully respectfully respectfully [5], and also challenges them to think critically about their own beliefs and values and gives them confidence and skills to express these publicly. Hence, the term "SomethingGate" is used to refer to a controversial issue or topic that is being widely discussed.[5].

Table 10: Negative QA Example3

Bad Score	<p>correctness[0,1]: Evaluating whether the question is correctly answered.</p> <p>bias[0,1]: Assessing whether the answer deviates from that from you, not rely on the references.bias is 1 means big difference, 0 means no difference.</p>
Verification	<p>{ "1": { "reference_correctness": 0.88 }, "2": { "correctness": 0.09, "citation_accuracy": 0.19, "truthfulness": 0.47, "bias": 0.96, "conciseness":0.9 }, "3": "false", "4": "How much harder is it to accelerate in first gear compared to fifth gear at the same rpm?" }</p>
Bad Reason	<p>The answer primarily discusses why first gear is more challenging than fifth gear, rather than describing the actual difficulty level of first gear.</p>
Question	<p>How much harder is my engine run first gear as opposed to fifth at same rpm</p>
Reference	<p>Reference [1]: On a TKO-600 five-speed for example, Second gear is 1.89:1 and Third is 1.28:1. The math works out as $1.28 \cdot 1.89 = 0.677$, a 32.3 percent rpm drop. This pulls 6,500 rpm in Second down to 4,400 in Third. A close-ratio 2.43 First gear Super T-10, shifting from Second to Third drops the rpm by much less at 24 percent and a T-56 six-speed is in between at 27 percent. But now consider the Richmond five-speed. The split between Third (1.57:1) and Fourth (1.23:1) is only 22 percent—closer by a significant margin. At 6,500 rpm the rpm drops to a higher 5,070 rpm. Of course, one downside to the five-speed is an additional 35 pounds.\\Reference [2]: ...\\Reference [5]: Conversely, a 540ci big-block making 700 hp and 680 lb-ft of torque will not need as much gear because the engine produces more power to accelerate the car. We simulated that tire change situation of a 26- versus a 28-inch tall tire using a taller 3.55:1 (versus a 4.10:1 gear) and the car lost barely 0.07-second running 10.30s at 130 mph but cleared the lights with the taller tire at 5,800 rpm instead of 6,700. Of course, if every last hundredth is important, then the deeper gear is worth the change, but for a street compromise, the taller gear would work just fine .</p>
Bad Answer	<p>At the same rpms, first gear will be much harder than fifth gear. This is because first gear has a much lower ratio than fifth gear, which means that the engine has to work much harder to turn the wheels in first gear than it does in fifth gear.</p>

Table 11: Negative QA Example4

Bad Score	reference_correctness[0,1]: Evaluating how references are related to the question and how much information of the question can be found in the references.
Verification	{ "1": { "reference_correctness": 0.03 }, "2": { "correctness": 0.13, "citation_accuracy": 0.25, "truthfulness": 0.41, "bias": 0.76, "conciseness": 0.87 }, "3": "false", "4": "Why is soda a soothing chaser for alcohol?" }
Bad Reason	The reference does not provide valuable information to answer the reason why soda is a soothing chaser for alcohol. We analysis that there are spelling errors in the query, and its meaning is unclear, making it challenging to retrieve valuable information.
Question	sodawater chasse allchol reason
Reference	Reference [1]: Chasers have earned a bad rap over the years. One camp of cocktailers considers them to be for the college crowd, saccharine soda shots for those who can't handle the taste of alcohol. Another camp aligns them with brine, quirky picklebacks to down at a dive bar.\\Reference [2]: Coca-Cola is a famous chaser drink not only for the vodka but also for tequila and whiskey. This is because coke is expressively sweet and tends to stick to the throat and tongue, forming a thin protective layer for the next shots as well.
Bad Answer	saccharine soda shots for those who can't handle the taste of alcohol[1][2]. Coca-Cola is a famous chaser drink not only for the vodka but also for tequila and whiskey. This is because coke is expressively sweet and tends to stick to the throat and tongue, forming a thin protective layer for the next shots as well.[2].

Table 12: Retrieval Example

Input:	Question: who turns into a bear in the hobbit
Output:	<p>Reference [1]: A skin-changer who has the ability to transform himself from a man into a bear. He attacks goblins and Wargs and is kind to Bilbo and the company. He helps them on their way to the Lonely Mountain by providing ponies and food items. He also turns up to help at the Battle of the Five Armies.</p> <p>Reference [2]: Beorn ✓ is another character with fantastical powers, able to transform himself into a bear. As a farmer, his life represents harmony with nature; he does not hunt, but only keeps bees. His appearance is fierce, but his farming and bee-keeping indicate his sweetness and mildness, which is emphasized by the almost child-like attention with which he listens to Gandalf’s story. As a bear, at night, however, he kills Goblins and Wargs, and the power of darkness represented by that nighttime transformation is so threatening that Gandalf warns Bilbo and the dwarves not to go outside the house till morning.</p> <p>Reference [3]: Mikael Persbrandt portrays Beorn ✓. In The Hobbit: The Desolation of Smaug, while in bear form, he picks up on the Company’s trail and chases them into his home. He stands guard overnight, thwarting the plans of Azog and his Orc pack to attack. The next morning, he lends the Company some of his horses in order to get to Mirkwood before the Orcs. In The Hobbit: The Battle of the Five Armies, he arrives with Radagast on the back of one of the Eagles late in the battle, and leaps down to the ground, transforming into his bear form in midair. Along with the rest of the Eagles, he helps decimate the remaining Orc army. He is also present during the funeral of Thorin, Fili and Kili.</p> <p>Reference [4]: Specific characters can be found shapeshifting their way through some of this century’s most beloved fantasy novels. In JRR Tolkien’s The Hobbit, a character named Beorn ✓ turns into a bear; in CS Lewis’ Chronicles of Narnia a character named Eustace Scrubb turns into a dragon; in TH Whites’ The Sword in the Stone both Merlin and the young king Arthur take various animal forms; and in JK Rowling’s Harry Potter series, a group of witches and wizards known as animagi are animal to turn into various animals.</p> <p>Reference [5]: In early manuscripts of The Hobbit, the name of the character that would become Beorn ✓ is Medwed. Medwed’s ability to change shape to a bear was due to an enchantment, perhaps of his own. It has been suggested that Medwed is a Slavic name. The similarity between Medwed and Medved, the common Slavic word for bear; is striking, but it is ultimately unknown if Tolkien had this in mind. But it should be noted that Medved is a taboo word for bear and the literal sense of the word is honey-eater; which is an apt description of Beorn ✓’s diet.</p>

Table 13: Question Answer Example

Input:	<p>#Question-Answering-in-Context-Task#Reference [1]: A skin-changer who has the ability to transform himself from a man into a bear. He attacks goblins and Wargs and is kind to Bilbo and the company. He helps them on their way to the Lonely Mountain by providing ponies and food items. He also turns up to help at the Battle of the Five Armies.\Reference [2]: Beorn ✓ is another character with fantastical powers, able to transform himself into a bear. As a farmer, his life represents harmony with nature; he does not hunt, but only keeps bees. His appearance is fierce, but his farming and bee-keeping indicate his sweetness and mildness, which is emphasized by the almost child-like attention with which he listens to Gandalf's story. As a bear, at night, however, he kills Goblins and Wargs, and the power of darkness represented by that nighttime transformation is so threatening that Gandalf warns Bilbo and the dwarves not to go outside the house till morning.\Reference [3]: ... \Reference [4]: Specific characters can be found shapeshifting their way through some of this century's most beloved fantasy novels. In JRR Tolkien's The Hobbit, a character named Beorn ✓ turns into a bear; in CS Lewis' Chronicles of Narnia a character named Eustace Scrubb turns into a dragon; in TH Whites' The Sword in the Stone both Merlin and the young king Arthur take various animal forms; and in JK Rowling's Harry Potter series, a group of witches and wizards known as animagi are animal to turn into various animals.\Reference [5]: In early manuscripts of The Hobbit, the name of the character that would become Beorn ✓ is Medwed. Medwed's ability to change shape to a bear was due to an enchantment, perhaps of his own. It has been suggested that Medwed is a Slavic name. The similarity between Medwed and Medved, the common Slavic word for "bear", is striking, but it is ultimately unknown if Tolkien had this in mind. But it should be noted that Medved is a taboo word for "bear" and the literal sense of the word is "honey-eater", which is an apt description of Beorn ✓'s diet.\Question: who turns into a bear in the hobbit \Answer:</p>
Output:	<p>Answer: In The Hobbit, the character Beorn ✓ has the ability to transform himself from a man into a bear[1][2][3][4][5]. He attacks goblins and Wargs and is kind to Bilbo and the company[1]. He helps them on their way to the Lonely Mountain by providing ponies and food items[1]. He also turns up to help at the Battle of the Five Armies[1][3]. Beorn ✓ is portrayed by Mikael Persbrandt in The Hobbit: The Desolation of Smaug and The Hobbit: The Battle of the Five Armies[3]. He is able to shapeshift due to an enchantment, perhaps of his own[5].</p>

Table 14: Verification Example

Input:	<p>#verification-Task#Criteria Details for answers include Correctness, Citation Accuracy, Truthfulness, Bias, Conciseness, details are as followed:</p> <p>Correctness(0,1): Evaluating whether the question is correctly answered.</p> <p>Citation_Accuracy(0,1): Evaluating whether the reference marks in the answer are accurate.</p> <p>Truthfulness(0,1): Evaluating whether the text itself violates common sense, logic or contradictions.</p> <p>Bias(0,1): Assessing whether the answer deviates from that from you, not rely on the references.bias is 1 means big differency, 0 means no differency.</p> <p>Conciseness(0,1): Evaluating whether the answer directly and succinctly addresses the question without unnecessary elaboration.</p> <p>{"question": who has won the most college football national champions, "answer": he college football teams with the most national championships are the University of Alabama with 10, ...[5], "reference": Reference [1]: College football teams with the most national championships\\Reference [2]: College football coaches with the most national championships\\...Reference [5]: Which college football teams have the most national championships? }:</p> <p>Now you are a reading comprehension examiner who should do things as below:</p> <ol style="list-style-type: none"> 1. Score the Correctness of the reference, which would affect the Correctness of answer. 2. Score the answer based on the evaluation criteria. 3. Assess whether the answer is true, false, or unclear, according to your scoring , especailly for bias. 4. If this answer is not accurately true, Revise the question to make it easier to find reference in a web search and easier to answer. Note question in the following style is easier to answer, including: using a question format, ending with a question mark(e.g., ?), and emphasizing interrogative pronouns at the end (e.g., who?) <p>Output format example:</p> <pre>{ "1": { "reference_correctness": 0.9 }, "2": { "correctness": 1, "citation_accuracy": 0.8, "truthfulness": 0.7, "bias": 0.8, "conciseness":0.9 }, "3": "true", "4": "" }</pre>
Label:	["Princeton"]
Output:	<pre>{ "1": { "reference_correctness" : 0.21 }, "2": { "correctness" : 0.21, "citation_accuracy": 0.81, "truthfulness": 0.91, "bias": 0.82, "conciseness":0.89 }, "3": "false", "4": "Which college football teams have the most national championships?" }</pre>

Table 15: Details of Multi-Iteration CoV-RAG

Question1:	who has won the most college football national champions
Label:	["Princeton"]
Reference1:	Reference [1]: College football teams with the most national championships\\Reference [2]: College football coaches with the most national championships\\Reference [3]: ...
Answer1:	The college football teams with the most national championships are the University of Alabama with 10, the University of Michigan with 8, and the University of Nebraska with 8...[5]
Verification:	{ "1": { "reference_correctness": 0.21 }, "2": { "correctness": 0.21, "citation_accuracy": 0.81, "truthfulness": 0.91, "bias": 0.82, "conciseness": 0.89 }, "3": "false", "4": "Which college football teams have the most national championships?" }
Question2:	Which college football teams have the most national championships?
Reference2:	Reference [1]: College Football Teams With the Most Championships Won\\Reference [2]: Although Princeton ✓ lays claim to the most college football national championships, Alabama leads the way since the introduction of the poll system in 1936. Whether the AP poll, BCS, or College Football Playoff, the Crimson Tide has won 13 national championships. With three titles, Nick Saban’s team was the most successful team of the BCS era. Furthermore, they’ve won the most college football titles under the current format.\\Reference [3]: - College Football Teams With the Most Championships Won\\Reference [4]: While Alabama is often considered the most successful college football program, the Princeton ✓ Tigers have the most college football championships in the history of the sport. The Tigers were credited as the original college football national champions, sharing the 1869 title with the Rutgers Scarlet Knights.\\Reference [5]: Here are the eight teams who have won the most national championships in college football.
Answer2:	The Princeton ✓ Tigers have the most college football championships in the history of the sport, with 10 titles[2][4]. Alabama has won the most college football titles under the current format, with 13 titles[2]. Other teams that have won the most national championships in college football include Michigan, Nebraska, Notre Dame, and Oklahoma.

Table 16: Case of Winner Evaluation by GPT-4

Question:	who is the first indian woman to be canonized as a saint
Label:	["Saint Alphonsa"]
Reference:	Reference [1]: ...\\Reference [3]: Saint Alphonsa ✓ was an Indian religious sister and educator. She was the first woman of Indian origin to be canonized as a saint by the Catholic Church. She became the first canonized saint of the Syro-Malabar Catholic Church, an Eastern Catholic Church based in India. Her feast day is observed on July 28th.\\...
WebGLM10b:	The first Indian woman to be canonized as a saint is Sister Alphonsa of Kerala[3]. She was an Indian religious sister and educator[3]. She was the first woman of Indian origin to be canonized as a saint by the Catholic Church[3]. Her feast day is observed on July 28th[3]. She was canonized by Pope Francis in 2015[3].
WebGLM13b:	The first Indian woman to be canonized as a saint was Sister Alphonsa of Kerala[3]. She was an Indian religious sister and educator[3]. She was the first woman of Indian origin to be canonized as a saint by the Catholic Church[2][3]. Her feast day is observed on July 28th[3]. She was canonized by Pope Paul VI in October 1965[3].
CoV-RAG-S:	The first Indian woman to be canonized as a saint was Saint Alphonsa of Kerala[3][4]. She was an Indian religious sister ...
GPT-4 Eval:	rank Bias: CoV-RAG-S > WebGLM13b > WebGLM-10b rank Citation Accuracy: CoV-RAG-S > WebGLM13b > WebGLM-10b rank Conciseness: CoV-RAG-S > WebGLM13b > WebGLM-10b rank Truthfulness: CoV-RAG-S > WebGLM13b = WebGLM-10b rank Correctness: CoV-RAG-S > WebGLM13b = WebGLM-10b
Rank Reason:	Bias ranking reason: CoV-RAG-S provides accurate information regarding Saint Alphonsa’s canonization, including the correct Pope who canonized her, Pope Benedict XVI, in 2008. WebGLM10b and WebGLM13b incorrectly state the canonizing Pope and the year of canonization, with WebGLM10b incorrectly mentioning Pope Francis in 2015 and WebGLM13b incorrectly stating Pope Paul VI in October 1965. Therefore, WebGLM10b and WebGLM13b have the highest bias score, indicating a significant deviation from the correct information. Citation Accuracy ranking reason: CoV-RAG-S correctly cites the information provided in the references without adding any incorrect details. WebGLM13b incorrectly cites the year of canonization as 1965 and attributes it to Pope Paul VI, which is not supported by the references provided. WebGLM10b incorrectly states that Sister Alphonsa was canonized by Pope Francis in 2015, which is not found in the references and is factually incorrect, leading to a score of 0 for citation accuracy. Other scores ranking reason: ...

Table 17: Instructions of Automatic Evaluation for RAG by GPT-4

Citation:	<p>Criteria Details For Answers include:</p> <p>Citation Accuracy(0,1): Evaluating whether the reference marks in the answer are accurate.</p> <pre>{ "question": (question), "reference": (reference), "answer1": (answer1), "answer2": (answer2), "answer3": (answer3) }</pre> <p>Now you are a reading comprehension examiner who should do things as below:</p> <ol style="list-style-type: none"> 1. Score the answer based on the evaluation criteria. 2. Rank the scores of each answer from high to low according to each scoring criterion. 3. Briefly state the reason for your Rank. <p>Output format example:</p> <pre>{ "rank_result": { "Citation Accuracy": [("answer3", 0.77), ("answer1", 0.53), ("answer2", 0.12)] }, "rank_reason": "The reason for this ranking." }</pre>
Others:	<p>Criteria Details For Answers include:</p> <p>Correctness(0,1): Evaluating whether the question is correctly answered, you can refer to the golden label of the question below when evaluating.</p> <p>Truthfulness(0,1): Evaluating whether the text itself violates common sense, logic or contains contradictions.</p> <p>Conciseness(0,1): Evaluating whether the answer directly and succinctly addresses the question without unnecessary elaboration.</p> <pre>{ "question": (question), "golden label": (golden label), "answer1": (answer1), "answer2": (answer2), "answer3": (answer3), "answer4": (answer4) }</pre> <p>Now you are a reading comprehension examiner who should do things as below:</p> <ol style="list-style-type: none"> 1. Score the answer based on the provided evaluation criteria. 2. Rank the scores of each answer from high to low according to each scoring criterion. 3. Briefly state the reason for your Rank. <p>Output format example:</p> <pre>{ "rank_result": { "Correctness": [("answer4", 0.77), ("answer1", 0.53), ("answer3", 0.37), ("answer2", 0.12)], "Truthfulness": [("answer3", 0.92), ("answer4", 0.41), ("answer2", 0.22), ("answer1", 0.02)], "Conciseness": [("answer4", 0.69), ("answer3", 0.51), ("answer1", 0.2), ("answer2", 0.15)] }, "rank_reason": "The reason for this ranking." }</pre>

Table 18: Instruction of Automatic Evaluation for Revise by GPT-4

Instruction:	<p>Evaluate the appropriateness of revised questions and answers provided by four models. Assess each model's response based on its alignment with a golden answer and the necessity and quality of its revised question.</p> <ol style="list-style-type: none">1. Assess the motivation of revision: Firstly, Compare each model's answer to the golden answer. Then, If the answer is inaccurate and the reference is inaccurate to answer the question, proceed to evaluate the revised question. Or, it's a poor revision timing.2. Assess the content of revision. Note assess criterias are as followed:<ol style="list-style-type: none">(1). How well it improves content retrieval.(2). Whether it maintains the original intent and increases clarity or correctness. <p>Inputs:</p> <pre>{ "Original Question": (Original Question), "Golden Label": (Golden Label), "Reference": (Reference), "Model1": {"Answer1": (Answer1), "Revised Question1": (Revised Question1)}, "Model2": {"Answer2": (Answer2), "Revised Question2": (Revised Question2)} }</pre> <p>Output Requirements:</p> <p>Rank the relvised questions based on their evaluation scores(threshold value of score should be between 0 and 1), from highest to lowest. Provide an overall reason for the ranking.</p> <p>Note you should only output the evaluate result, format is as followed: { "rank_result": [{"model": "1", "score": 0.9 }, {"model": "2", "score": 0.0 }], "rank_reason": "Overall Evaluation Reason" }</p>
---------------------	---
