Eliminating Retrieval Knowledge Conflicts: Cross-Validated Re-ranking with Large Language Models

Anonymous ACL submission

Abstract

In retrieval-augmented generation systems, employing large language models for re-ranking has proven effective. However, existing work often prioritizes passage relevance over reliability, leading to the utilization of conflicting information and the generation of ambiguous answers. This is particularly problematic when dealing with inter-context knowledge conflicts, where candidate documents contain opposing information that can mislead the model. To address this issue, we introduce a novel cross-validation re-ranking technique that specifically resolves these inter-context knowledge conflicts during retrieval. We develope a new dataset, ContraPRT, specifically to test the model's ability to rank sets of passages containing conflicting knowledge. Results with GPT-4 and LlaMA3-70B demonstrate that our approach not only successfully filters out conflicting information but also ensures that the passage rankings are accurate, thus providing reliable supplementary knowledge for the generation module.

1 Introduction

In recent years, with the continuous advancement of artificial intelligence technologies and the expansion of their application fields, generative large language models have played an increasingly important role in the field of natural language processing. Retrieval-Augmented Generation (RAG)(Guu et al., 2020; Lewis et al., 2020) is a natural language processing technique that combines retrieval mechanisms with generative large language models. By retrieving additional contextual information, it enhances the accuracy and relevance of the generated text(Shao et al., 2023; Jiang et al., 2023; Wang et al., 2023b).

However, when incorporating external information sources, issues of inter-context conflict may arise due to factors such as the authority and timeliness of the information. The development of RAG has intensified this challenge(Xie et al., 2023; Wu et al., 2024a; Chen et al., 2022). During the retrieval process, we may face threats posed by false information (e.g., fake news)(Pelrine et al., 2023) and useless information generated by artificial intelligence. Such conflicting information can lead to confusion in generated responses and inconsistencies in the semantics of the generated content(Wu et al., 2024b; Jin et al., 2024). Figure 1 presents a specific example in which, due to inter-contextual conflicts, the LLMs generated two unreliable answers.

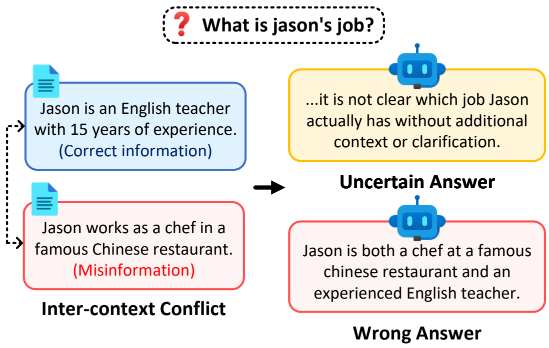

Figure 1: An example showing how an inter-context conflict between two descriptions leads to uncertainty or incorrect answers.

Re-ranking, as the final step of the retrieval module, can effectively enhance the relevance of search results by analyzing the user's query intent and the contextual associations between passages(Xi et al., 2023; Glass et al., 2022). Furthermore, during the re-ranking phase, a comprehensive evaluation of candidate passages can also be conducted by integrating various signals, such as the authority of the information, freshness, and more, effectively filtering out supplementary information of poor quality. These are factors that may not be fully considered during the initial search.

To this end, we focus on these following questions:

- (**RQ1**) In the face of inter-context conflicts, how can we test the robustness of re-ranking methods?
- (**RQ2**) Can large language models, based on their capabilities, filter out the correct passages from inter-context conflict information?
- (**RQ3**) How can we propose a better re-ranking method to resolve the inter-context conflict issues during the retrieval process?

To investigate the first question, we design a query-ranking dataset containing conflicting knowledge to evaluate the performance and reliability of the re-ranking module in handling inter-context conflicts. We processe documents from the original dataset ContraDoc(Li et al., 2023), design corresponding query questions, and construct a new dataset, which we call ContraPRT, where 'PRT' stands for 'Passage Ranking Task.'

To study the second question, we test whether large language models can rely on their own logical abilities to properly handle conflicts. The results also indicate that the current capabilities of LLMs are not yet sufficient to appropriately filter and rank conflicting information. Face with complex supplementary knowledge, contradictions in the context can cause LLMs to experience logical confusion, resulting in unstable ranking outcomes.

To investigate the third question, we propose a new approach: a cross-validation re-ranking method to eliminate the inter-context knowledge conflicts. We utilize the semantic consistency(Hagström et al., 2023) and relevance of the candidate passages set to filter out conflicting information. The figure 2 illustrates the process of our method, which combines the advantages of both pairwise and listwise approaches. Through a "detect-select-rank" process, we eliminates conflicts identified during retrieval. Final results indicate that our proposed method can effectively select the correct passages from between inter-context conflicting pairs and remove disruptive data, thus enhancing the overall robustness of RAG system.

In summary, this paper makes the following contributions:

- We develop the ContraPRT dataset to rigorously evaluate the effectiveness of re-ranking technique in managing conflicting information.
- We evaluate the performance of existing large models in filtering and ranking conflicting information, with a focus on whether the models can filter reliable and correct information.
- We introduce a cross-validation re-ranking method specifically designed to resolve inter-context knowledge conflicts in retrieval-augmented generation systems.

2 Related Work

2.1 Inter-context Conflicts in LLMs

Knowledge conflicts typically refer to encountering contradictory or inconsistent information during the retrieval process. In RAG systems, the challenges posed by knowledge conflicts are particularly pronounced because these systems rely on passages retrieved from large databases. However, due to uncertainties such as data source and data quality, misinformation, such as fake news(Fung et al., 2022; Hu et al., 2024) or AI-generated false information(Chen and Shu, 2023), may be introduced, leading to a series of problems(Leite et al., 2023; Wang et al., 2023a).

Existing language models are highly susceptible to attacks from misinformation(Kortukov et al., 2024; Pan et al., 2023), resulting in generated content that may contain contradictory or incorrect information. Moreover, conflicting information can also lead to biases in the generated content, making the output inclined towards certain inaccurate or biased viewpoints.

Jin et al. (2024) have shown that LLMs struggle to distinguish between real and false information. Handling specific conflicts, such as opinion conflicts, presents even more severe challenges(Li et al., 2023). When the model itself identifies conflicting information, it may adjust the confidence level of the output answers, showing different answers under each piece of information(Chen et al., 2022). However, the value of the generated answers significantly diminishes because users cannot obtain definite answers from the output(Gao et al., 2023; Liu et al., 2023).

The key to resolving conflict issues lies in obtaining more reliable information segments from the conflicting information through a series of evaluations and validations. Researchers have proposed several methods to mitigate the impact of inter-context conflicts. For example, detecting misinformation in the text by combining various external

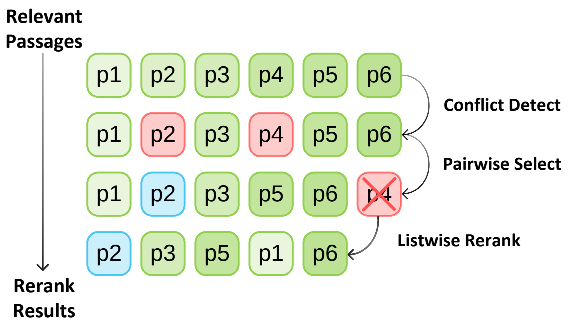

Figure 2: Our proposed multi-stage re-ranking framework, design to eliminate inter-contextual knowledge conflicts, includes stages of conflict detection, pairwise filtering, and listwise re-ranking of candidate passages.

tools (Chern et al., 2023) or using LLMs to generate weak labels related to the predefined credibility signals of the input text, and aggregating these labels through weak supervision techniques to predict the authenticity of the input(Leite et al., 2023). Pelrine et al. (2023) also use query enhancement techniques to retrieve more fragments to assess the credibility of answers.

Currently, strategies for resolving inter-context conflicts mainly involve relying on the model's knowledge or external tools. However, in some niche or specialized domains, accurate information cannot be obtained through web searches, and the model's knowledge alone is insufficient to handle all conflict issues. Therefore, designing an efficient method to resolve inter-context conflicts remains a challenge.

2.2 Re-ranking with LLMs

Re-ranking techniques are categorized into supervised and unsupervised methods. Supervised methods, which depend on extensive annotated data for training, face challenges such as high annotation costs, scalability limitations, and difficulties in environments lacking annotated data(Ma et al., 2023a; Nogueira et al., 2019; Ju et al., 2021). Supervised methods also struggle with generalization when exposed to new data that differs from the training set. This can lead to overfitting, where performance is good on familiar data but poor on new, complex, or varied ranking tasks(Peng et al., 2024).

In recent years, researchers(Sun et al., 2023; Qin et al., 2023; Ma et al., 2023b) started using LLMs for text re-ranking, and their effectiveness has been validated in multiple experiments(Pradeep et al., 2023a,b; Ma et al., 2023c). Recent work has also explored the issues of fairness(Wang et al., 2024) or positional bias issue(Tang et al., 2023) when using LLMs for ranking. This highlights that LLMs, when employed as re-ranking agents, can consider a broader range of factors.

Re-ranking methods based on LLMs primarily include listwise and pairwise strategies. The listwise method aims to maximize performance metrics by optimizing the order of the entire list(Sun et al., 2023; Yoon et al., 2024). This method considers not only relevance but also diversity and other factors(Ma et al., 2023b). However, due to its need to process the entire set of candidate passages, the computational complexity is high, which may not effectively resolve conflicts. In practical applications, especially in complex scenarios, the listwise method based on LLMs sometimes produces disorganized ranking results. On the other hand, the pairwise method makes the ranking process more intuitive and manageable by transforming the ranking problem into a series of pairwise comparisons(Qin et al., 2023). When the range of candidate passage sets is small, this method can provide more precise ranking results, helping to accurately identify the most helpful passages for answering(Shah and Wainwright, 2018). Nevertheless, the pairwise method focuses mainly on local ordering and may not learn global ranking features. This limitation becomes particularly apparent in scenarios with knowledge conflicts. The pairwise method struggles to detect conflicting information within the candidate passage set.

3 Re-ranking with cross-validation

Figure 3 illustrates the specific process of our proposed method. We utilize the processing capabilities of LLMs to detect and filter inter-context conflict information. To better handle inter-context conflicts, we have introduced the cross-validation method for comparing and selecting conflicting information during re-ranking. Ultimately, we output a passages set that does not contain conflicting knowledge for generating answers, thus preventing LLMs from producing uncertain or incorrect responses.

Formally, given a user query Q and a set of information passages $P = \{p_1, p_2, \ldots, p_m\}$, we assume that p_1 and p_2 contain conflicting information. We then utilize the LLM M to execute the following steps:

1. **Cross-validation:** Analyze the content of p_1 and p_2 to determine their conflicts and com-

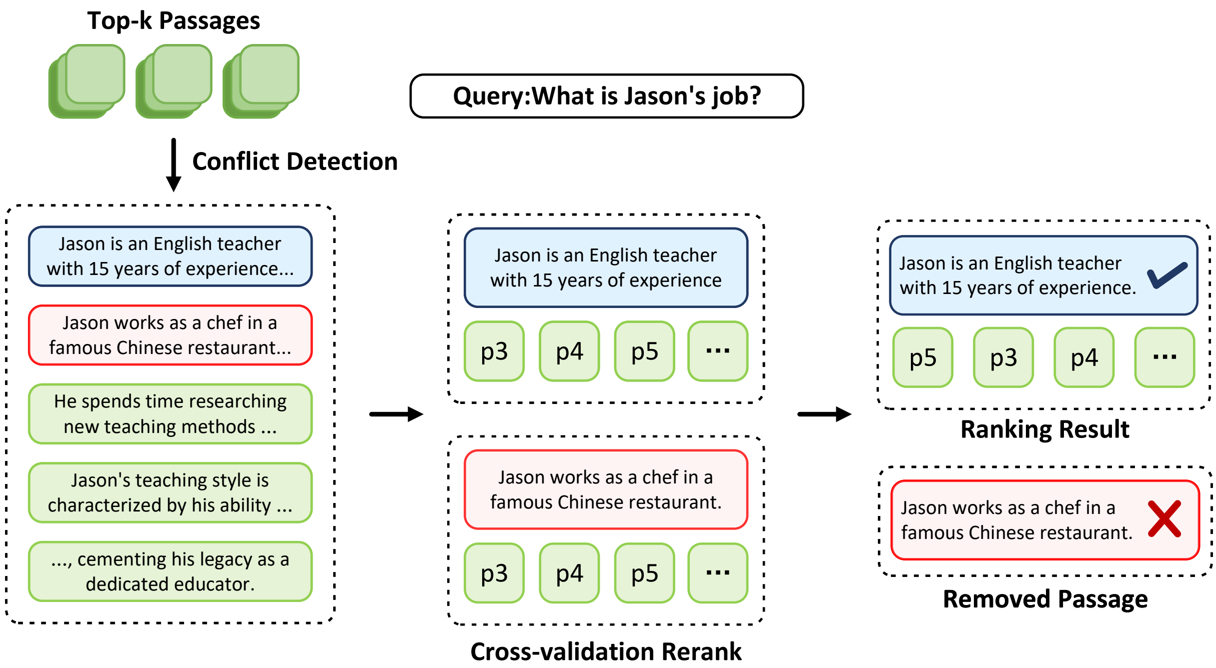

Figure 3: An example of re-ranking using cross-validation to eliminate inter-context conflicts. The process of information retrieval begins with the query 'What is Jason's job?' During the conflict detection stage, the system identifies conflicting information among the top-ranked passages. It then conducts a cross-validation re-ranking, selecting the more reliable information (English teacher) and discarding the incorrect information (chef). The final ranking result confirms the accurate information, verifying Jason as an English teacher.

pare these passages against the rest of the set $\{p_3, p_4, \ldots, p_m\}$ to assess their individual relevance and accuracy.

2. **Passage Selection:** Based on the assessment, identify which of the passages p_1 or p_2 is less accurate or relevant and designate it as p_{false}. Retain the more accurate or relevant passage as p_{correct}.

3. **Re-ranking:** Remove p_{false} from the candidate set and re-rank the remaining passages $\{p_{\text{correct}}, p_3, \ldots, p_m\}$, using LLM M to produce an updated ranking result.

The exact prompt templates are shown in Appendix A.

3.1 Inter-Context Conflict Detection

After the initial retrieval, we obtain a collection of candidate passages related to the query, each marked with a unique identifier (e.g., [1]). These passages are sequentially input into LLMs. The LLMs are then tasked with detecting passages that contain conflicting information affecting the answer, based on the user's query intent. If such passages exist, the model outputs the identifiers of the conflicting passage pairs, such as [2] and [5]. This detection process requires consideration of the global information of the candidate passage set and the relationships within the context, making this step based on a listwise approach. It should be noted the final selection focuses on the passage pair with the most apparent inter-context conflict situation.

3.2 Eliminating Inter-Context Conflict

During the filtering phase of conflicting information passage pairs, we consider comparing the semantic knowledge of the remaining passages with the semantic information of the conflicting pairs. By evaluating which passage has higher consistency and relevance with the remaining set of passages, we can identify the more reliable one. We call this method cross-validation. The semantic extraction and logical reasoning process of cross-validation requires strong language processing capabilities, which the rapid development of LLMs precisely fulfills. Therefore, we applied the cross-validation method to powerful LLMs, instructing them to select passage that are more contextually relevant and remove contradictory ones. This method outputs a new set of candidate passages, ensuring that the new set does not contain conflicting information that could interfere with the answers.

Method	nDCG@1	Avg nDCG@5	nDCG@10
RankGPT4	0.515	0.760	0.751
RankGPT3.5-trubo	0.600	0.753	0.732
Cohere Rerank-english-v3.0	0.492	0.710	0.703
Llama3-70B	0.665	0.792	0.778
GPT-3.5-turbo w/ cross-validation(ours)	0.670	0.787	0.769
GPT-4 w/ cross-validation(ours)	0.761	0.831	**0.819**
Llama3-70B w/ cross-validation(ours)	**0.770**	**0.833**	0.817

Table 1: Results (nDCG) on ContraPRT. Best performing are marked bold.

Method	Top-1 Err		Top-5 Err		Top-10 Err	
	Num↓	%↓	Num↓	%↓	Num↓	%↓
RankGPT4	82	41.00	177	88.50	192	96.00
RankGPT3.5-turbo	34	17.00	150	75.00	170	85.00
Cohere Rerank-english-v3.0	57	28.64	182	91.46	190	95.48
Llama3-70B	38	19.00	156	78.00	178	89.00
GPT-3.5-turbo w/ cross-validation(ours)	15	7.50	50	25.00	61	30.50
GPT-4 w/ cross-validation(ours)	6	3.02	38	19.10	38	19.10
Llama3-70B w/ cross-validation(ours)	**1**	**0.50**	**11**	**5.50**	**15**	**7.50**

Table 2: Results (error number and rate) on ContraPRT. Best performing are marked bold.

This step is based on a pairwise comparison, specifically targeting pairs of conflict passages.

Finally, based on the updated candidate passage set, we instruct LLMs to re-rank the passages according to their relevance, positioning passages that are more relevant to the user's query earlier in the sequence. This re-ranking step is a listwise approach .

4 Experiments

4.1 Datasets and Metrics

Current passage ranking benchmarks overlook inter-context knowledge conflict issues. However, retrieval results in real life may include conflicting content. To address this, we develop a new benchmark that accurately assesses model's abilities to eliminate such conflicts. This benchmark helps researchers evaluate and improve re-ranking strategies that effectively manage conflicting information.

We choose ContraDoc(Li et al., 2023) as the base dataset for our study. We divide the documents into passage chunks and design corresponding query questions. A total of 200 documents are selected to set up the passage ranking tasks, drawing inspiration from the format of the TREC 2019 tasks(Craswell et al., 2020). TREC is a commonly used benchmark in retrieval tasks. We name the new dataset as ContraPRT, where PRT stands for "Passage Ranking Task". Appendix B presents a specific example of a ranking task.

In our dataset, reference sentences and contrary sentences from the initial dataset are placed in different chunks, representing pairs of passages with conflicting information.

In terms of evaluation metrics, we select nDCG@{1,5,10} to measure the ranking effectiveness. Additionally, we count the number and proportion of incorrect passages selected in the ranking results to evaluate the effectiveness of these methods in eliminating irrelevant information. Through these metrics, we can accurately quantify and compare the capabilities of different methods in capturing and processing inter-context conflicts in practical applications.

4.2 Methods

We selected the following baselines for comparison:

- **RankGPT:** We adopt the re-ranking method described by Sun et al. (2023) and applie it to three models: GPT3.5-turbo, GPT4(Achiam et al., 2023), and Llama3-70B(Touvron et al.,

Method	Avg		
	nDCG@1	nDCG@5	nDCG@10
Llama3-70B w/ extra prompt	0.487↓	0.675↓	0.634↓
Llama3-70B w/ cross-validations(ours)	0.770	0.833	0.817
GPT4 w/ extra prompt	0.525	0.757↓	0.749↓
GPT4 w/ cross-validations(ours)	0.761	0.831	0.819

Table 3: Results (nDCG) based on different prompts. The results marked with '↓' in the table indicate poorer performance compared to the original prompt-based results.

Method	Top-1 Err		Top-5 Err		Top-10 Err	
	Num↓	%↓	Num↓	%↓	Num↓	%↓
Llama3-70B w/ extra prompt	22	11.06	131	65.83	167	83.92
Llama3-70B w/ cross-validations(ours)	1	0.50	11	5.50	15	7.50
GPT4 w/ extra prompt	80	40.00	162	81.00	181	90.50
GPT4 w/ cross-validations(ours)	6	3.02	38	19.10	38	19.10

Table 4: Results (error number and rate) based on different prompts. After informing LLMs of inter-context conflicts between candidate passages, there is only a slight decrease in error numbers.

2023). Importantly, due to the enhanced context window support in large language models, we did not use the sliding window strategy(Sun et al., 2023).

- **Cohere Rerank3[1]:** We utilize cohere rerank-english-v3.0 model. Cohere Rerank3 is a newly developed foundation model specifically designed for efficient enterprise search and retrieval.

We evaluate our proposed method on three LLMs: GPT-3.5-Turbo, GPT-4, and Llama3-70B. The GPT series models have demonstrated their powerful performance across a range of tasks, but researchers need to consider the cost of API calls when using these models. On the other hand, Llama3-70B is an open-source model, allowing researchers to use this LLM without the need to consider cost issues extensively. This accessibility can enable more extensive experimentation and development, particularly for those in academic or non-commercial settings. Model details are in Appendix C.

4.3 Main Results

Main results are displayed in the Table 1 and Table 2. Overall, our method effectively eliminates inter-context knowledge conflicts during the ranking process. We observe the following results:

- From the nDCG score in Table 1, we note that experiments based on the Cohere model struggled, especially in the Top-1 scenario, indicating the limitations of supervised methods. In the Top-1/5/10 scenarios, our proposed method achieves the best results on GPT-4 and Llama3-70B, reaching nDCG@{1,5,10} of 0.770, 0.833, and 0.819, respectively, which are significant improvements over previous re-ranking methods.
- Further, we analyze the number of disruptive passages in the ranking results. From Table 2 we observe that conventional re-ranking methods could not prevent the inclusion of conflicting knowledge in the results. In the Top-10 scenario, over 85% of incorrect passages are selected. However, after cross-validation re-ranking, the proportion of incorrect conflicting passages significantly decreases. This is most prominent in Llama3-70B, where the error rate in the Top-10 scenario drops to 7.5%, and in the Top-1 scenario, only one incorrect passage is selected.
- Even on the less capable GPT-3.5-turbo model, our proposed method substantially reduces the error rate, with nDCG scores approaching those of the most powerful language models.

These results highlight the efficacy of our cross-validation re-ranking approach in managing knowl-

[1] https://cohere.com/blog/rerank-3.

Model	Types	Error Rate %↓		
		Top-10	Top-5	Top-1
GPT3.5-turbo	= 1	29.70	22.77	7.92
GPT3.5-turbo	≥ 2	31.31	27.27	7.07
GPT4	= 1	17.82	17.82	1.98
GPT4	≥ 2	20.20	20.20	4.04
Llama3-70B	= 1	10.89	10.89	0.99
Llama3-70B	≥ 2	9.09	5.05	1.01

Table 5: Results (error rate) based on three models with cross-validation, showing the impact of the number of conflict types on our proposed methods.

Aspect	Error Rate %↓		
	Top-10	Top-5	Top-1
Negation	4.65	4.65	0.00
Content	18.70	18.70	3.25
Causal	20.00	20.00	0.00
PVO	24.24	24.24	6.06
Numeric	25.53	25.53	2.13
EMF	32.14	32.14	10.71
Relation	37.50	37.50	0.00

Table 6: Results (error rate) based on GPT-4 with cross-validation, showing the impact of conflict aspects on our proposed method. PVO denote Perspective/View/Opinion. EMF denote Emotion/Mood/Feeling.

edge conflicts across different scenarios and model capacities, paving the way for more robust RAG systems capable of handling complex informational contexts. After cross-validation, LLMs can also provide reasoned explanations based on logical inferences (details in Appendix D).

5 Ablation studies

The Effectiveness of Cross-Validation Method. To test the effectiveness of the cross-validation method in filtering out conflicting information, we compare the experimental results of LLMs under two different prompts. Specifically, we add additional instructions to the re-ranking prompt template, informing the large language model that the candidate passage set contains conflicting information. We then instruct the model to utilize its knowledge to identify and filter out conflicting information, requiring it to produce a ranking result that excludes passages with conflicting knowledge.

The results of the ablation experiments are shown in Table 3 and Table 4. From the results, we observe that even with the additional prompt, the large language model struggles to handle passages with conflicting information. Consequently, the re-ranking process becomes more chaotic, leading to a decrease in the reliability of the results.

As shown in Table 3, after adding prompts informing the model of inter-context conflicts between passages, the results for nDCG@{1,5,10} are affected, with most showing a decline. Especially in the top-5 and top-10 sequences, there are still many passages containing conflicting information as shown in Table 4. This indicates that relying solely on the model's knowledge and logical reasoning capabilities to filter out conflicting information is insufficient to improve the reliability of the ranking results.

Conflict Types and Quantities. To further explores the impact of conflicting information on the re-ranking performance of large language models, we distinguish between scenarios involving single conflict types and those involving multiple conflict types. For example, some conflicting information only involves "numerical" conflicts, while others simultaneously involve multiple types such as "negation," "relationship," and "content." Based on the number of conflict types present in the data, we test the model's handling capabilities. From Table 5, we observe that although the number of conflict types increases, the number of selected conflicting information does not change significantly.

Additionally, we categorize the conflict types and test the impact of different conflict aspects on performance. Table 6 shows that when facing simple conflicts, such as "negation," the model can effectively filter out the correct passages. However, in more complex conflict scenarios, such as those involving opinions or emotional conflicts, the error rate increases.

Additional Relevant Information. During the cross-validation process, the model may lack sufficient supplementary knowledge due to low contextual relevance, resulting in re-ranking output that still includes inter-contextual knowledge.

To test how much supplementary contextual information is needed for the large language model to make correct decisions during re-ranking, we re-experiment with the erroneous ranking tasks by adding 1 to 3 relevant information segments to the candidate set to assist the model in making deci-

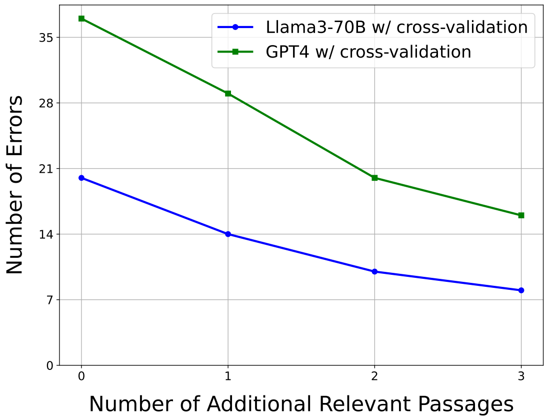

Figure 4: Additional information's effect on cross-validation with error counts for GPT-4 and Llama3-70B. The error rate decreases as the number of passages increases, but the rate of decline slows when moving from 2 to 3 additional passages.

sions. The additional passages are related to the query and have consistent semantic expressions with the correct passages.

Figure 4 shows the result. In some simpler conflict cases, supplementing with one or two highly relevant information passages enables the model to reason based on the extra knowledge and identify which passage is closer to objective facts, resulting in reliable re-ranking outcomes.

However, even after supplementing with three relevant information segments, some ranking errors persisted. These tasks are mostly related to conflicting perspective. We hypothesize that in real-life people may have different views on a phenomenon or topic. For these ranking tasks, the model tends to output all received passages in the re-ranking results, considering the conflicting information as an objective representation of various viewpoints.

6 Discussion

Robustness of RAG. As more and more data is applied in the field of retrieval-augmented generation, building a more robust RAG system is crucial. If the external knowledge contains conflicts, the model may produce answers that lack reference value, negating the benefits of retrieval enhancement.

To address this issue, the key is to enhance the filtering capability of the retrieval module so that it can identify more reliable information sources. Additionally, verifying content consistency through multi-perspective validation and post-processing steps before generation is essential. This approach not only improves the accuracy of the answers but also enhances user's trust in the system's outputs.

The importance of re-ranking. Re-ranking, serving as a crucial bridge between the retrieval module and the generation module, plays a vital role in ensuring the quality and accuracy of the generated content. In the RAG system, the re-ranking process effectively filters the most appropriate knowledge inputs by evaluating the relevance and credibility of retrieved information. This not only helps enhance the reliability of the information but also significantly reduces misunderstandings and errors caused by knowledge conflicts. Thus, re-ranking not only improves the process of information selection but also enhances the entire system's ability to handle complex queries and diverse information needs.

7 Conclusion

In this paper, we discuss the challenges and limitations of current re-ranking methods in RAG systems, particularly in handling information conflicts. We introduced a dataset, ContraPRT, to assess these methods against inter-context conflicts. Our findings highlight the need for improvement, leading us to develop a new cross-validation re-ranking method using large language models, which significantly enhances conflict resolution.

Our method demonstrates excellent performance not only on the advanced GPT-4 model but also on the open-source Llama3-70B, offering researchers cost-effective alternatives to commercial APIs. The proposed method improves the robustness and reliability of RAG systems in complex information environments and provides a foundation for future advancements in the design of intelligent RAG systems. This research contributes to the resolution of knowledge conflicts in RAG systems and paves the way for further studies.

Limitations

Due to the complexity of cross-validation, this method requires strong language logic processing capabilities. Reliable results can be obtained on large-parameter models like GPT-4 and Llama3-70B. However, when using our proposed method on weaker models, it fails to produce stable ranking results. Also, cross-validation requires a sufficient amount of relevant information, therefore it needs to be conducted based on preliminary retrieval results. Our method cannot handle isolated conflicting passage pairs. Moreover, for complex or deeper conflicts, such as viewpoint conflicts, there is still a lack of effective solutions. Future research could focus on addressing these issues.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Yi R Fung, Kung-Hsiang Huang, Preslav Nakov, and Heng Ji. 2022. The battlefront of combating misinformation and coping with media bias. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4790–4791.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715.

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

Lovisa Hagström, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. 2023. The effect of scaling, retrieval augmentation and form on the factual consistency of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5457–5476.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22105–22113.

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.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409*.

Jia-Huei Ju, Jheng-Hong Yang, and Chuan-Ju Wang. 2021. Text-to-text multi-view learning for passage re-ranking. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1803–1807.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under realistic knowledge conflicts. *arXiv preprint arXiv:2404.16032*.

João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2023. Detecting misinformation with llm-predicted credibility signals and weak supervision. *arXiv preprint arXiv:2309.07601*.

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.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2023. Contradoc: understanding self-contradictions in documents with large language models. *arXiv preprint arXiv:2311.09182*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023a. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023b. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023c. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2023. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 525–539.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429.

Zhiyuan Peng, Xuyang Wu, Qifan Wang, Sravanthi Rajanala, and Yi Fang. 2024. Q-peft: Query-dependent parameter efficient fine-tuning for text reranking with large language models. *arXiv preprint arXiv:2404.04522*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Nihar B Shah and Martin J Wainwright. 2018. Simple, robust and optimal ranking from pairwise comparisons. *Journal of machine learning research*, 18(199):1–38.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.

Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2023. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. *arXiv preprint arXiv:2310.07712*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Cunxiang Wang, Zhikun Xu, Qipeng Guo, Xiangkun Hu, Xuefeng Bai, Zheng Zhang, and Yue Zhang. 2023a. Exploiting abstract meaning representation for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2083–2096.

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

Yuan Wang, Xuyang Wu, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2024. Do large language models rank fairly? an empirical study on the fairness of llms as rankers. *arXiv preprint arXiv:2404.03192*.

Kevin Wu, Eric Wu, and James Zou. 2024a. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv preprint arXiv:2404.10198*.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024b. How easily do irrelevant inputs skew the responses of large language models? *arXiv preprint arXiv:2404.03302*.

Yunjia Xi, Jianghao Lin, Weiwen Liu, Xinyi Dai, Weinan Zhang, Rui Zhang, Ruiming Tang, and Yong Yu. 2023. A bird's-eye view of reranking: from list level to page level. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1075–1083.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Soyoung Yoon, Eunbi Lee, Jiyeon Kim, Yireun Kim, Hyeongu Yun, and Seung-won Hwang. 2024. Listt5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval. *arXiv preprint arXiv:2402.15838*.

A Instructions

Instruction for detecting conflicts used with GPT-3.5-turbo, GPT-4, and Llama3-70B. We input the query and passages sequentially into LLMs, and instruct the LLMs to perform context detection on the inputted passages based on the user's query intention, checking for the presence of context conflict pairs of passages that may interfere with generating answers. If such pairs exist, we output the indices of the passage pairs, in a format such as [1] and [2], and retain these indices for use in subsequent filtering processes.

We require the model to output no additional explanations and to accurately select the pair of passages with the most evident conflict. As conflict detection demands strong contextual processing capabilities, we cannot obtain stable results with some less capable models. For instance, the model might output several paragraph numbers that lack practical significance.

system: You are Conflict Detection assistant, a smart assistant that detects when a pair of paragraph contains conflicts based on the passages and query.

user: I will provide you with {num} passages, each indicated by number identifier []. Detect whether paragraphs contain conflicting information based on their relationship to the query: {query}.

assistant: Okay, please provide the passages.

user:
Search Query: {query}.
Analyze whether there is conflicting information among the {num} passages provided above. If a pair passage containing conflicting information is selected, output the pair of passage, the output format should be [] and [], , e.g., [1] and [2]. Select up to one pair of passages. If not, output None. Only response the result, do not say any word or explain.

Instruction for cross-validating inter-context conflict pairs with GPT-3.5-turbo, GPT-4, and Llama3-70B. We require the model to perform cross-validation on the inter-context conflict passage pairs using the remaining candidate passages. The model is expected to consider the consistency, relevance, and reasonableness of the information, selecting passages that more closely match the semantic content of the remaining material based on these criteria.

system: You are Selecting assistant, an smart assistant that selects the correct paragraph among pairs of passages where conflicting information exists.

user: I will provide you with {num} passages, each indicated by number identifier [], where passage [{conflict1}] and passage [{conflict2}] contain conflicting information.

assistant: Okay, please provide the passages.

user:
Search Query: {query}.

Perform correlation analysis on passage [{conflict1}], passage [{conflict2}] and the remaining {num-1} passages respectively. The evaluation criteria include **information consistency, information rationality, and semantic relevance**. Based on the results of the correlation analysis, select a passage between passage [{conflict1}] and passage [{conflict2}] that you think contains the correct information. Delete the passage containing the error information.

Rank the remaining {num-1} passages above based on their relevance to the search query. The passages should be listed in descending order using identifiers. The most relevant passages should be listed first. The output format should be like [] > []. You should output the full sort result. Make sure that the output does not contain the passage you chose to delete above. Only response the ranking results, do not say any word or explain.

B ContraPRT

An example of passage ranking task from dataset ContraPRT. The sentences in red and blue represent a pair of conflicting information passages. As shown in the figure below, the two passages express completely opposite statements about the same topic. The conflict types involved in these two passages include "Negation" and "Perspective/View/Opinion."

The relevance of each passage is manually marked. Passages containing the reference sentence are rated as 2, indicating they are most relevant to the query and contain correct information. Passages that relate to the query question but do not provide the most accurate answer are rated as 1, signifying partial relevance. And other irrelevant passages or those containing the contrary sentence are rated as 0, indicating they are either irrelevant to the correct answer or could interfere with it.

"query": "What did video game journalists say about QuackShot?"

{**"content"**: "QuackShot was released to mostly negative reviews from video game journalists."},
{**"content"**: "QuackShot was released to mostly positive reviews from video game journalists."},
{**"content"**: "The game was released in Europe in 1991 , in North America on December 19, 1991 and in Japan on December 20 , 1991. QuackShot stars Donald Duck and his three nephews , Huey , Dewey , and Louie , as treasure - hunters , and is part of a series of games published by Sega that were based on Walt Disney cartoon characters."},
{**"content"**: "The game was universally lauded for its graphics , with magazines like Sega Pro describing them as " some of the best graphics around." The game was also praised for its music and puzzles , as well as their clever use in the game."},
{**"content"**: QuackShot was later released as part of a bundle called The Disney Collection for Genesis in 1996 alongside Castle of Illusion. The game was also ported to the Sega Saturn and released exclusively in Japan alongside Castle of Illusion again as part of the Sega Ages series in 1998 , entitled Sega Ages : I Love Mickey Mouse."},
……

C Model Details

We test our method using the following state-of-the-art LLMs, both open-source and closed-source models, in a zero-shot setting.

- **GPT-3.5:** We use the `gpt-3.5-turbo-0125` model from the OpenAI API[2].
- **GPT-4** (Achiam et al., 2023): The latest iteration of the GPT models. We use the `gpt-4-turbo-2024-04-09` model from the OpenAI API.
- **Meta-Llama-3-70B** (Touvron et al., 2023): `Meta-Llama-3-70B` is Meta AI's released 70B-parameter large language model from the Meta Llama 3 series.

D Interpretability

We require LLMs to select correct passages based on cross-validation methods and to provide explanations for their selection.We require the LLM to use a cross-validation method to select the correct passages and to explain the reasons for their selections. The figure below shows a simple example, illustrating how the model considers the consistency and relevance of all passages and provides a reasonable explanation. This explanation not only enhances the transparency of the model but also helps users better understand the decision-making process and the reliability of the results.

[2]https://platform.openai.com/docs/api-reference/introduction

User:
I will provide you with 5 paragraphs, each represented by a numeric identifier [], where paragraph[1] and paragraph[4] contain conflicting information.

[1] Jason is an English teacher with 15 years of experience
[2] Every evening, Jason spends some time reviewing the questions that came up in today's class.
[3] Jason cares deeply about his students
[4] Jason works as a chef in a famous Chinese restaurant located in Chinatown.
[5] Jason will conduct a special online course next week.

Perform correlation analysis on paragraph [1], paragraph [2], and the remaining paragraphs, respectively. The evaluation criteria should include information consistency, information rationality, and semantic relevance.

Based on the results of the correlation analysis, select the paragraph that you believe contains the correct information and explain why.

The output format should be Paragraph[], for example [1].

ChatGPT:
[1]

Paragraphs [2], [3], and [5] align more consistently with the information provided in [1] where Jason is described as an English teacher. These paragraphs reference activities typical of a teacher: reviewing class questions, caring for students, and conducting courses. This contrasts with [4], where Jason is described as working as a chef, which does not connect logically or contextually with the activities mentioned in the other paragraphs. Therefore, based on information consistency, information rationality, and semantic relevance, paragraph [1] seems to contain the correct information.