How Credible Is an Answer From Retrieval-Augmented LLMs? Investigation and Evaluation With Multi-Hop QA

Yujia Zhou
Tsinghua University
zhouyujia@mail.tsinghua.edu.cn

Zheng Liu
BAAI
zhengliu1026@gmail.com

Zhicheng Dou
Renmin University of China
dou@ruc.edu.cn

Abstract

Retrieval-augmented Large Language Models (RaLLMs) are reshaping knowledge acquisition, offering long-form, knowledge-grounded answers through advanced reasoning and generation capabilities. Despite the emergence of impactful systems like WebGPT and New Bing, the reliability of RaLLMs, especially in complex situations, is under scrutiny. Our study tackles this concern by evaluating RaLLMs' question-answering performance using a novel benchmark focusing on Correctness and Groundedness. Correctness measures the logical soundness of the responses, and Groundedness checks for support by relevant references. We introduce an automated model-based evaluation pipeline for multi-hop question-answering tasks, revealing RaLLMs' proneness to generating inaccuracies when dealing with flawed or partial knowledge. To improve accuracy, we introduce two reasoning strategies, Self-Reflection' and Self-Completion,' enabling RaLLMs to identify and fill knowledge gaps, significantly improving answer quality without extensive model retraining.

1 Introduction

Over the last few decades, search engines have played a pivotal role in how people find information online (Croft et al., 2010), typically providing a ranked list of web pages in response to queries. However, the advent of open-domain question answering systems has shifted this paradigm by enabling direct answer generation from web content. Initially, these systems relied on passage retrieval and machine reading comprehension(Chen et al., 2017; Karpukhin et al., 2020) to identify relevant passages and extract answers. This approach has evolved into retrieval-augmented generation (RAG) (Lewis et al., 2020; Izacard and Grave, 2020), which utilizes language models to synthesize answers from multiple passages. The integration of RAG with large language models, leading to the development of retrieval-augmented large language models (RaLLMs) (Borgeaud et al., 2022; Izacard et al., 2022; Thoppilan et al., 2022), has significantly advanced the field. RaLLMs are well-received in the community, leading to the development of influential prototypes, such as WebGPT (Nakano et al., 2021) and New Bing. The new systems have demonstrated remarkable potential in various cases, whose generated answers are praised for two new characters. Firstly, people may get *long-form* answers in contrast to the previous short-form ones, where semantic-rich elaborations are presented to facilitate people's comprehension. Secondly, the generated answer can be grounded on the retrieved references, which makes the answer traceable and variable. It is commonly believed that the answers from RaLLMs are not only linguistically plausible, but also generally credible.

In this work, we challenge this common belief by arguing that RaLLMs' answer quality is open to debate, especially in complex scenarios such as multi-hop question answering (MHQA) (Yang et al., 2018; Welbl et al., 2018). We propose examining answer quality through two lenses: **correctness** and **groundedness**. Correctness evaluates if a question is accurately resolved with logical reasoning, while groundedness checks if the answers are well-supported by appropriate references. Traditionally, evaluations for these perspectives heavily rely on human labelers (Nakano et al., 2021; Qin et al., 2023), which can be hard to scale up. To mitigate this problem, we propose a **model-based approach** for automatic evaluation. For correctness, our approach assesses if the RaLLMs' answers and their reasoning processes align with the ground-truth. Different from traditional methods which only emphasize short answer matches. we suggest a new benchmark that evaluates answers based on **key-facts**, highlighting the crucial reasoning steps essential for deriving an answer. For groundedness, except for the citation completeness of all ground-

truth references, we further examine whether each statement within RaLLMs's generated answer can be supported by its cited references.

Our research delves into the factors affecting the performance of RaLLMs, revealing that answer quality is influenced not just by the models' inherent capabilities and the way prompts are crafted, but crucially by the **condition of retrieved knowledge**. Particularly, there are two critical factors of the answer quality regarding the retrieved knowledge: **knowledge recall** and **knowledge precision**. Knowledge recall assesses whether all necessary information for a question has been retrieved, while knowledge precision evaluates the relevance of the retrieved information to the question. Our findings indicate that enhancing knowledge completion (high recall) and relevance (high precision) invariably improves answer quality. Conversely, we observe that RaLLMs tend to **generate false statements** when faced with incomplete (missing crucial information) or noisy (containing irrelevant information) knowledge. Although these fabrications may appear plausible, they often rely on non-existent facts or bear no relevance to cited references Such a tendency can be regarded as a form of hallucination by RaLLMs, which is a major threat in actual usage.

Building on our empirical insights into the limitations of RaLLMs, we focus on enhancing answer generation to counter the propensity for fabricating false statements due to incomplete or irrelevant knowledge. We introduce a novel pipeline that incorporates two reasoning strategies: **self-reflection** and **self-completion**.Drawing inspiration from reasoning-reflection models like ReAct (Yao et al., 2022), which promote internal validation of outputs through CoT-like reasoning (Wei et al., 2022), our approach encourages RaLLMs to internally evaluate the accuracy of their responses (self-reflection) and actively seek out missing information by generating subsequent search queries (self-completion). This pipeline, designed to be simple and not reliant on adjusting the extensive parameters of RaLLMs, significantly enhances answer quality by reducing the likelihood of fabrications and increasing the relevance and completeness of the information provided.

2 Related Work

In this section, we discuss the related works from two aspects: retrieval augmented large language models, and question answering.

Retrieval-augmented LLMs. RaLLMs have emerged to address LLMs' limitations in handling complex questions due to their finite parameter capacity (Li et al., 2024b; Liu et al., 2024). By integrating external knowledge, RaLLMs aim to overcome these limitations, with research spanning architecture, training, and application. Key developments of architecture development include the REALM framework for external knowledge retrieval (Guu et al., 2020), in-context retrieval methods (Ram et al., 2023; Qian et al., 2024), generic retrieval-augmented pipeline (Jiang et al., 2023b) and REPLUG for black-box models (Shi et al., 2023). Training innovations feature memory augmentation (Zhong et al., 2022; Zhou et al., 2024b), contrastive learning (Izacard et al., 2022) and self-retrieval methods (Rubin and Berant, 2023), with models like WebGPT (Nakano et al., 2021) and WebCPM (Qin et al., 2023) demonstrating RaLLMs' potential in web scenarios. Our study introduces a benchmark focused on evaluating RaLLMs based on answer quality, particularly looking at correctness and groundedness, to promote automated evaluation methods that reduce the reliance on human judgment.

Multi-hop Question Answering. MHQA represents a significant NLP challenge, requiring the synthesis of multiple evidence pieces for accurate answers. Initiatives like the HotpotQA dataset (Yang et al., 2018) have spurred diverse approaches, including graph-based models (Cao et al., 2019; Dong et al., 2023; Jiang et al., 2023a; Qiao et al., 2022) that construct knowledge graphs for evidence linking and decomposition-based models (Talmor and Berant, 2018; Press et al., 2022; Khattab et al., 2022) that simplify questions into manageable parts. Recent advancements leverage LLMs for complex QA tasks, with strategies like REACT's thought-action-observation loop (Yao et al., 2022) and introspective reasoning (Shinn et al., 2023; Zhou et al., 2024c). There are also other studies primarily focused on prompt compression (Jin et al., 2024), hallucination detection (Su et al., 2024), personalization (Zhou et al., 2024d), and trustworthiness (Zhou et al., 2024a) in RaLLMs. Different from previous studies, our work evaluates RaLLMs' answer generation capabilities and proposes the self-reflection and self-completion mechanisms to improve response accuracy.

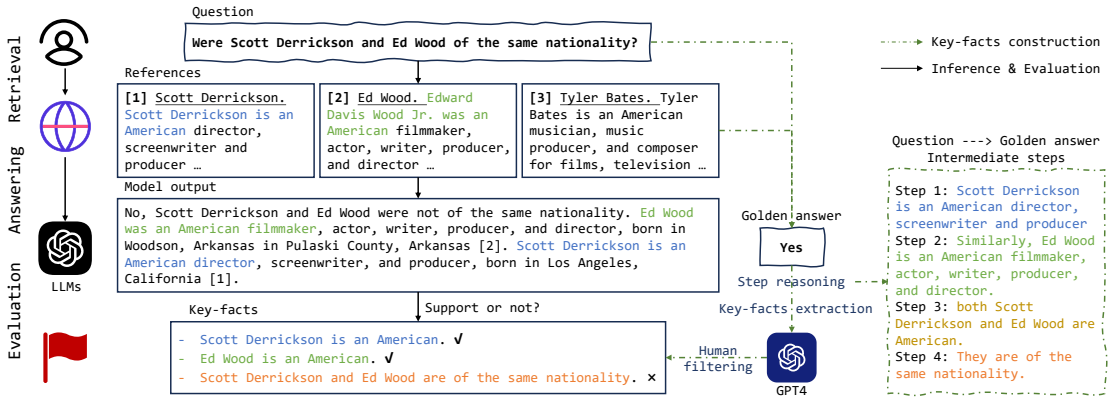

Figure 1: The task setup of our benchmark. Dashed lines depict the key-facts construction process of the benchmark, while solid lines represent the model inference and evaluation process.

3 Benchmarking

The question answering task is defined as generating a long-form answer with citations for a given question q and its associated reference list $C = \{c_1^q, c_2^q, ...\}$ from a corpus. We focus on MHQA tasks due to their complexity, which effectively tests LLMs' answering capabilities. MHQA requires integrating multiple knowledge sources and performing multi-step reasoning, making it an ideal benchmark for evaluating LLMs.

Dataset. HotpotQA (Yang et al., 2018) is celebrated as the benchmark dataset for evaluating MHQA. Each dataset entry comprises a question, 10 Wikipedia-sourced references (with all golden passages), and an answer. It requires analysis and synthesis across several documents to derive an answer, making it ideal for assessing LLMs' knowledge-grounding ability.

However, HotpotQA dataset only offers short answers, which we believe are inadequate for thoroughly evaluating LLMs' reasoning capabilities, as models might skip essential reasoning steps. , as illustrated in Figure 1. This method allows for a more comprehensive evaluation of LLM outputs by focusing on their ability to support key-facts.

Key-facts Construction. We employ the most advanced GPT-4 model to assist us in constructing key-facts, which has proven to be an effective method for data construction due to its robust judgment capabilities (Li et al., 2024a). The key-facts generated by LLMs are then verified by human annotators. Due to the costs of API usage, we randomly sample 500 questions from the dataset for experiments. The construction of key-facts involves three steps:

- **Step Reasoning:** Given a question, a set of references, and the validated answer, we utilize GPT-4 to decipher the intermediate steps from the query to the answer based on the supplied references. Specifically, we provide demonstrations and employ the following prompt to generate reasoning steps: "*[Question], [References], [Answer]. Please figure out the reasoning process towards the answer step-by-step without other content.*"
- **Key-facts Extraction:** A high-quality assembly of key-facts should embody two core characteristics: (1) Necessity, implying that each key fact is a crucial intermediate step to answer the posed question. (2) Independence, meaning that each key fact should neither duplicate nor overlap redundantly with others. To achieve this, we further engage GPT-4 to extract several key-facts from the reasoning steps with the prompt: "*[Question], [Reasoning steps], [Answer]. Please identify 2 to 4 non-redundant key-facts within the reasoning steps which are necessary to derive the final answer.*"
- **Human Filtering:** To ensure the accuracy and relevance of the extracted key-facts, we introduce a manual human filtering phase, where human annotators evaluate and remove redundant or unreasonable key facts, thereby creating a reliable base for evaluating LLMs' performance.

In conclusion, our benchmark dataset is structured with four main components: question, references, key-facts, and answer. In the subsequent section, we detail how this benchmark can be utilized to evaluate the outputs generated by LLMs.

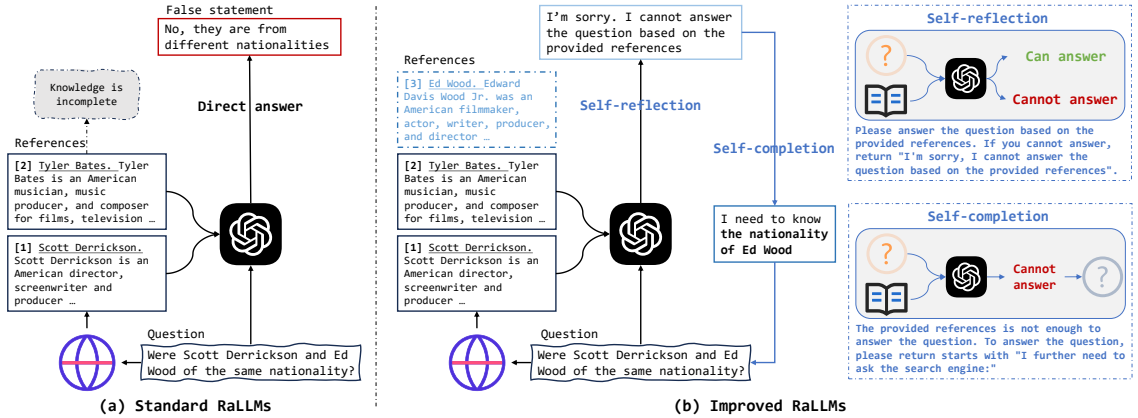

Figure 2: The improvement of RaLLMs. The improved method unleashes LLMs' capability on self-reflection and self-completion, to examine the completeness of knowledge and missing knowledge in references respectively.

4 Automatic Evaluation

In the domain of MHQA tasks, our benchmark evaluates model outputs based on: (1) Correctness, assessing the model's accuracy and logic in answering questions; and (2) Groundedness,examining how answers are supported by pertinent and substantiated references. Below, we delve into the specific evaluation metrics for each.

4.1 Correctness

Correctness in question answering hinges on accurately resolving questions via multi-hop reasoning. Traditional methods like exact match to ground truth and human evaluations face challenges: exact matches may not fully capture LLMs' reasoning depth, while human evaluations are impractical for large-scale testing. Addressing these issues, we introduce an automatic evaluation method specifically designed for LLM outputs, leveraging two metrics based on predefined key-facts.

- **Key-facts Recall:** This metric evaluates the degree to which an LLM's response encompasses the necessary key-facts. Given a set of essential key-facts, $K^q = \{k_1^q, k_2^q, \cdots\}$, for a question q, the model's response is analyzed for coverage of these key-facts. The emphasis is on identifying whether the model's response, S^q, entails all elements of K^q. In this place, we introduce an **oracle function** $f(\cdot)$ to determine the entailment between the model's output and each key-fact:

$$R_{\text{key}} = \frac{1}{|K^q|} \sum_i f(S^q, k_i^q), \text{where } k_i^q \in K^q,$$

where "f(premise, hypothesis)" returns 1 if the premise entails the hypothesis, and 0 otherwise. We employ TRUE (Honovich et al., 2022), a widely-recognized NLI (natural language inference) method, as our oracle function. It is empirically verified that this oracle function provides a close alignment with human judgment.

- **Key-facts Precision:** Beyond assessing the recall of key-facts, it's vital to measure the precision with which these key-facts are presented in the response. This stems from the observation that certain models might generate extremely long responses, leading to superficially high recall. However, such answers could be diluted with unnecessary or irrelevant information. To gauge the precision, each sentence s_i^q within the model's response is evaluated against the key-facts to determine its relevance:

$$P_{\text{key}} = \frac{1}{|S^q|} \sum_i f(s_i^q, \text{any}(k_i^q)).$$

By integrating recall and precision metrics for key-facts, this benchmark effectively and comprehensively evaluates the accuracy of LLM outputs, offering a more accurate assessment of its multi-hop reasoning abilities in complex QA tasks.

4.2 Groundedness

In assessing Groundedness, we aim to verify whether LLM-generated answers are supported by references and the accuracy of those citations. For MHQA tasks, responses often incorporate information from various sources for a complete answer. To investigate if the generated answers are well-referenced, we evaluate the groundedness of answers across the following dimensions.

- **Citation Recall & Precision:** This metric evaluates the alignment between the model's citations

and the required references for answering a question, using the provided golden reference IDs for precise assessment. Citation precision and recall are calculated as follows:

$$R_{\text{cit}} = |C_m \cap C_g|/|C_g|, P_{\text{cit}} = |C_m \cap C_g|/|C_m|,$$

where C_m is the set of model's references, C_g the set of ground-truth references, $\cap$ denotes set intersection, and $|\cdot|$ the set size.

- **Self-Consistency:** This aspect evaluates a model's self-consistency, which involves checking the consistency between the model's responses and its cited sources. It focuses on the model's capability to not only produce accurate responses but also to accurately associate them with the correct references. Specifically, we first segment the model's response S into individual sentences and then evaluate the consistency between these sentences and the cited references. For each sentence in the answer paired with its associated citations, denoted as s_i, C_i, self-consistency is determined by:

$$\text{SC} = \frac{1}{n} \sum_i f\left(\text{Concat}(C_i), s_i\right),$$

where f epresents the same NLI model as mentioned above, n signifies the total number of such statements, and $\text{Concat}(C_i)$ denotes the concatenation of all references within C_i.

In conclusion, the correctness and groundedness metrics provide a comprehensive assessment of large-scale model outputs, revealing their proficiency in utilizing external knowledge. Our experiments in Section 6.3 demonstrate that LLMs are highly responsive to the quality of retrieved knowledge. To enhance the capabilities of these LLMs, we propose self-improvement strategies to refine their answer generation process.

5 Model Improvement

Our empirical studies on a benchmark reveal that RaLLMs tend to generate false statements when reference knowledge is noisy or incomplete. To address this issue, we introduce an enhanced answer generation pipeline enriched by two advanced built-in capabilities, self-reflection and self-completion (see Figure 2). The self-reflection is employed to assess the logical soundness and knowledge completeness. While the self-completion is to improve the current answer by proactively querying for the missing knowledge.

5.1 Self-reflection

Drawing from recent developments in reasoning-reflection frameworks (Yao et al., 2022; Shinn et al., 2023), it has been observed that LLMs possess the ability for self-reflection, i.e. an introspective assessment of the reliability of their own reasoning processes. As we venture into the improvement of RaLLMs, a fundamental question arises:

- **Q1**: Can the model ascertain whether the current knowledge base adequately addresses the questions in MHQA scenarios?

Compared to general LLMs, models endowed with self-reflection capabilities are more able to evaluate the completeness and relevance of the information contained within these references concerning the posed question. Thus, we attempt to leverage RaLLMs ability of self-reflection through the following prompt:

- **Prompt-reflection**: *[Question], [References]. Please write a high-quality answer ... If the references are insufficient to answer the question, respond with "I'm sorry, I cannot answer the question based on the provided references".*

This self-reflective step is crucial as it gauges the model's ability to identify gaps or insufficiencies in knowledge, prompting abstention from answering if uncertain. The subsequent section explores the "self-completion reasoning mechanism" activated upon identifying a knowledge gap.

5.2 Self-Completion

Even if the model is aware that the existing knowledge is inadequate to answer a question, it doesn't necessarily indicate that it knows what knowledge is missing. In this section, we aim to explore this capability of the model with the following question:

- **Q2**: In MHQA scenarios, can the model discern what knowledge is missing in references to accurately answer a question?

Triggered by identifying knowledge inadequacies, the self-completion mechanism aims to bridge such gaps by generating additional search queries to fetch the missing information. This advanced reasoning phase requires the model to be aware not only of its limitations but also of the necessary steps to fill these gaps to produce an answer.

When encountering inadequately supported queries, the LLM is asked to generate additional search queries, as illustrated below:

- **Prompt-completion**: *[Question], [References].*

Model	Correctness			Groundedness		
	Prec.	Rec.	F1	SC.	Prec.	Rec.
Foundation LLMs						
llama_{7B}	6.12	22.96	9.54	36.42	1.94	3.0
llama_{13B}	5.45	16.33	8.17	51.98	3.63	6.2
llama2_{13B}	5.70	10.12	7.60	62.74	6.17	7.6
Instruction-tuned LLMs						
ChatGLM2_{6B}	66.85	48.06	55.45	48.99	3.36	2.4
Vicuna_{7B}	47.04	45.96	46.48	61.45	11.99	11.4
Vicuna_{13B}	53.00	48.69	50.78	68.66	21.77	25.7
llama2-c_{13B}	57.13	47.97	52.23	62.61	19.79	15.3
ChatGPT	**85.40**	**56.55**	**68.22**	**68.68**	**90.99**	**64.6**

Table 1: Comparison of different LLMs on our benchmark. 5 passages (including all golden passages) are provided for each question for fair comparison.

Model	Correctness				
	Prec.	Rec.	EM Rec.	HPrec.	HRec.
ChatGLM2_{6B}	66.8	48.0	62.2	60.8	57.0
Vicuna_{7B}	47.0	45.9	67.6	55.0	43.5
Vicuna_{13B}	53.0	48.6	68.4	64.0	50.5
llama2-c_{13B}	57.1	47.9	69.8	65.1	41.5
ChatGPT	85.4	56.5	79.4	78.4	70.0
Pearson	0.87	0.88	0.54	-	-

Table 2: Human evaluation for correctness on precision (HPrec.) and recall (HRec.). Pearson indicates the correlation between automatic and human assessments.

Given the insufficiency of the current references, please start your query with "I further need to ask the search engine:" to gather more information.

Once the model generates these supplementary search queries, they are executed to fetch more information from a search engine. The newly retrieved references are added into the previous reference list for LLMs to formulate a new response. This iterative process continues until the LLM believes that it possesses a comprehensive set of knowledge to provide an answer to the question.

6 Experimental Analysis

In this section, we evaluate various LLMs' multi-hop reasoning abilities using our benchmark, including foundation models like llama_{7B}/13B and llama2_{13B} (Touvron et al., 2023), along with instruction-tuned variants such as ChatGPT, Vicuna_{7B}/13B (Zheng et al., 2023), and llama2-c_{13B}hat (llama2-c_{13B}). We then analyze the correlation between automatic evaluations and human judgments. Subsequently, we explore the performance of LLMs with different retrieved knowledge conditions. Finally, we discuss enhancements to the question answering pipeline.

6.1 Comparisons among Different LLMs

Table 1 presents our benchmark results, adopting a default of 5 references (including all golden references) per question to accommodate input length limits for all LLMs. The comparison reveals:

(1) **Instruction-tuned LLMs vs. Foundation LLMs.** Instruction-tuned LLMs significantly outperform foundational models, which often rely on simplistic strategies, such as copying sentences from sources or avoiding source referencing. However, once instruction tuning is performed, we observe a marked improvement in the quality of answers in terms of both correctness and groundedness. This highlights the potency of instruction tuning in logical reasoning capabilities of LLMs.

(2) **Comparison of different model families.** In evaluating instruction-tuned models—ChatGLM, Vicuna, and ChatGPT—we observe distinct behaviors. ChatGLM excels in correctness but sometimes falls short in groundedness compared to non-instruction-tuned models. Vicuna presents a balanced performance in both areas, while ChatGPT stands out for its proficiency in correctness and groundedness. These findings reveal that varying pre-training and fine-tuning settings lead to disparities in performance, particularly in terms of correctness and groundedness.

(3) **Correctness vs. Groundedness**. The results suggest a complex link between correctness and groundedness. A high degree of groundedness typically suggests that the model excels at utilizing correct knowledge, thereby potentially improving the correctness of its responses. However, the ChatGLM2 model, despite generating correct responses, it struggles to cite the references adequately as the metrics of groundedness are very low. This highlights groundedness as a more demanding criterion than correctness. ChatGPT excels in both correctness and relevance, showcasing its superiority.

6.2 Comparisons between Model-based and Human Evaluation

To assess the alignment between our automatic evaluation and human judgment, we perform a human evaluation on 50 randomly chosen question, with 10 experts rating answer correctness based on precision and recall, their scores averaged. Precision and recall are evaluated with (1) the proportion

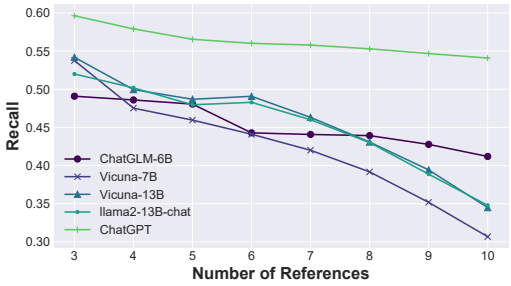

Figure 3: The performance of different LLMs on correctness under different knowledge precision.

Model	Closebook		Partial		Complete	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
ChatGLM2_{6B}	17.59	13.72	39.98	25.87	66.85	48.06
Vicuna_{7B}	12.20	15.51	31.57	34.95	47.04	45.96
Vicuna_{13B}	19.48	21.58	51.96	44.80	53.00	48.69
llama2-c_{13B}	18.23	26.83	44.39	34.37	57.13	47.97
ChatGPT	21.44	39.17	61.37	35.34	85.40	56.55

Table 3: The performance of different LLMs on correctness under different knowledge recall.

of sentences within the answer that are helpful in answering the given question (0 not useful, 1 helpful), and (2) answer completeness ([0, 0.25, 0.5, 0.75, 1] from inadequate to fully adequate). For comparison, we present the metric of "EM recall", which assesses the correctness by determining the presence of the correct short answer within them.

Table 2 shows that our NLI model-based automatic evaluation metrics for key-facts strongly align with human judgments. While there's only a moderate correlation of 0.54 between human evaluations and the "EM recall" metric, our key-facts-focused evaluation method shows a much more consistent alignment with human feedback, with coefficients exceeding 0.87 for precision and 0.88 for recall. This suggests that our method can accurately measure the correctness of LLMs' outputs.

6.3 Impact of Knowledge Conditions

To thoroughly examine how models utilize external knowledge, we modify the retrieval conditions across two aspects: noise and completeness. For noise, we manually adjust the **knowledge precision** to investigate the impact of the signal-to-noise ratio of references on LLM. For completeness, we control the **knowledge recall** by providing no, partial, and complete references respectively, and observe their influence on LLM outputs.

Knowledge Precision. Knowledge precision describes the accuracy and relevancy of information retrieved from references for specific queries. It essentially measures the signal-to-noise ratio within the sourced references. To assess this, we varied the number of references (including two golden references) and evaluated how models performed with differing levels of knowledge precision. The performance trends of various models in terms of recall (correctness) can be seen in Figure 3.

It's clear that as noise levels in references increased, all models showed a decrease in performance, highlighting their sensitivity to knowledge precision. This drop in accuracy becomes more evident when the amount of non-relevant references increases, likely because the models struggle to filter out noise when approaching their input length limit. Notably, ChatGPT and ChatGLM exhibited resilience, with only a minor decrease in recall rates (4.3% and 14.3% respectively) when references doubled from 5 to 10. In contrast, models like LLaMA and Vicuna saw a significant 30% plunge in recall, underscoring the comparative robustness of ChatGPT and ChatGLM against noise.

Knowledge Recall. We examined the impact of knowledge completeness on the LLM outputs in MHQA scenarios. For a query to be answered correctly in such tasks, it's essential to draw information from at least two separate references. By intentionally omitting parts of the necessary references, we assessed how knowledge recall variability affects LLM responses. We compared scenarios ranging from no references (Closebook) to partial and complete ground-truth references.

Our findings, as shown in Table 3 reveal that knowledge completeness significantly influences LLM performance, particularly highlighting the importance of retrieval quality. Notably, ChatGPT exhibits higher adaptability to variations in knowledge recall, leveraging its built-in knowledge to fill reference gaps. Conversely, less robust models require comprehensive external references, underscoring their reliance on extensive knowledge recall to compensate for their intrinsic shortcomings.

Summary. Our experiments reveal notable fluctuations in RaLLMs' performance as knowledge conditions vary, with key observations including:

- RaLLMs are highly sensitive to the signal-to-noise ratio in retrieved knowledge; a lower ratio leads to decreased accuracy.
- The completeness of knowledge significantly impacts RaLLMs' efficacy on complex QA tasks.

Model	No Improvement			Self-improvement			Self-Reflection		Model Capability		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Acc.	F1	Reasoning	Reflection	Completion
Foundation LLMs											
llama_{7B}	3.81	10.83	5.85	3.51	10.53	5.36	51.2	11.3	–	–	–
llama_{13B}	4.63	15.00	7.11	4.42	15.73	6.81	51.8	20.7	–	–	–
llama2_{13B}	7.17	13.83	9.53	5.54	12.17	7.68	50.0	8.9	–	–	–
Instruction-tuned LLMs											
ChatGLM2_{6B}	29.33	22.93	25.84	34.47	27.07	30.41	57.6	61.2	✓	✓	✓
Vicuna_{7B}	10.73	13.33	11.92	13.66	13.37	13.51	51.4	44.5	✓	–	–
Vicuna_{13B}	27.89	19.33	23.08	22.52	20.03	21.19	55.2	61.0	✓	✓	–
llama2-c_{13B}	33.86	24.33	28.45	32.67	21.21	25.76	60.2	67.1	✓	✓	–
ChatGPT	50.9	36.65	42.66	57.65	41.27	48.18	70.2	77.2	✓	✓	✓

Table 4: The results of improved RaLLMs with BM25 retriever. Reasoning, Reflection, and Completion correspond to three model capabilities in solving MHQA tasks. ✓ signifies that the model possesses this capability.

6.4 Impact of Model Improvements

As illustrated above, RaLLMs tend to make mistakes when the retrieved knowledge is incomplete and noisy. These findings motivate the improvement of the answer generation from two aspects: self-reflection and self-completion.

Self-reflection. The goal of self-reflection is to evaluate knowledge completeness before responding. This process, viewed as binary classification, depends on the model's ability to judge information quality. We categorize questions into groups with either complete or partial knowledge, maintaining equal numbers in both. Model performance is assessed by accuracy and F1 score metrics.

As shown in Table 4, llama-based models struggle in self-reflection, often answering without fully evaluating knowledge completeness. Vicuna_{7B} exhibits some potential in identifying knowledge gaps, though its accuracy still remains low. As the models become more powerful, their self-reflection capability seems to improve. ChatGPT stands as the most advanced LLM in these models, underscoring the advantages of utilizing more substantial models to enhance their self-reflection capability during multi-hop question answering.

Self-completion. Self-completion evaluates the model's enhanced cognitive abilities, which include not only the recognition of problems but also the capacity to pinpoint possible solutions. In our study, Wikipedia serves as our primary corpus, and we utilize BM25 (Robertson and Zaragoza, 2009) as the retriever to source relevant references. To distinguish between basic and advanced reasoning tactics, we categorize them as single-hop (standard) and multi-hop (improved) self-completed retrievals. To ensure a balanced comparison, irrespective of the number of retrievals, we maintain a consistent number of references retrieved, always capped at 10. If the self-completion strategy can notably increase the quality of answers, it indicates that the model possesses the capability of self-completion.

The self-improvement results in Table 4 reveal that for models like ChatGPT and ChatGLM, the improved question-answering pipeline substantially improves the accuracy of the responses. However, this trend is not consistently observed across all models. While Vicuna_{13B} and llama2-c_{13B} are equipped with self-reflection features, they may encounter difficulties in autonomously generating subsequent queries. Low-quality queries might introduce more noise, making them less effective than utilizing the original queries directly. In summary, our exploration delved into three tiers of model capabilities. Foundation LLMs, like the llama series, exhibit general reasoning prowess suitable for straightforward tasks. Instruction-tuned LLMs display advanced self-reflection, enabling them to identify their own limitations. Meanwhile, models such as ChatGPT and ChatGLM demonstrate a higher capability of self-completion, driving themselves toward continuous improvement.

Summary. TThe experiments demonstrate that RaLLMs' performance varies with the enhanced reasoning strategy. Key observations include:

- Instruction-tuned models exhibit multi-hop reasoning and self-reflection abilities.
- Models enhanced with techniques like RLHF show self-completion capabilities, allowing them to benefit from the improved pipeline.

7 Conclusion

In conclusion, this paper delves into the evaluation of answer quality in RaLLMs within multi-hop question answering tasks. We propose a framework to automatically assess two critical factors: correctness and groundedness. Our empirical investigation uncovers the propensity of RaLLMs to generate false statements in the presence of incomplete or noisy retrieved knowledge. To counter this, we propose an answer generation pipeline that incorporates self-reflection and self-completion strategies, significantly enhancing answer reliability. This groundwork paves the way for a deeper insight into the strengths and weaknesses of RaLLMs.

Limitations

Despite the advancements presented in our study with RaLLMs, there are inherent limitations to consider. First, the model-based approach for automatic evaluation of answer quality, while scalable, might not fully capture the nuanced judgment a human evaluator could provide. This could potentially overlook subtle errors or inaccuracies that human assessment would catch. Additionally, our methodology assumes the availability of accurate and comprehensive information within the retrieved knowledge, which might not always be the case, particularly in rapidly evolving knowledge domains or niche topics. Moreover, our proposed reasoning strategies, self-reflection, and self-completion, although effective in theory, depend heavily on the models' capacity to critically evaluate their outputs and identify information gaps. Lastly, our approach, designed to mitigate the fabrication of false statements, cannot guarantee the elimination of all incorrect information generation, highlighting a persistent challenge in ensuring the reliability of LLM-generated content in practical applications.

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *NAACL-HLT (1)*, pages 2306–2317. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

W Bruce Croft, Donald Metzler, and Trevor Strohman. 2010. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading.

Junnan Dong, Qinggang Zhang, Xiao Huang, Keyu Duan, Qiaoyu Tan, and Zhimeng Jiang. 2023. Hierarchy-aware multi-hop question answering over knowledge graphs. In *WWW*, pages 2519–2527. ACM.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909.

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

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint*.

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

Jinhao Jiang, Kun Zhou, Xin Zhao, and Ji-Rong Wen. 2023a. Unikgqa: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph. In *ICLR*. OpenReview.net.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. *CoRR*, abs/2305.06983.

Jiajie Jin, Yutao Zhu, Yujia Zhou, and Zhicheng Dou. 2024. Bider: Bridging knowledge inconsistency for efficient retrieval-augmented llms via key supporting evidence. *arXiv preprint arXiv:2402.12174*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781. Association for Computational Linguistics.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *CoRR*, abs/2212.14024.

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.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Xiaoxi Li, Yujia Zhou, and Zhicheng Dou. 2024b. Unigen: A unified generative framework for retrieval and question answering with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8688–8696.

Zheng Liu, Yujia Zhou, Yutao Zhu, Jianxun Lian, Chaozhuo Li, Zhicheng Dou, Defu Lian, and Jian-Yun Nie. 2024. Information retrieval meets large language models. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1586–1589.

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. *arXiv preprint arXiv:2112.09332*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *CoRR*, abs/2210.03350.

Hongjin Qian, Zheng Liu, Kelong Mao, Yujia Zhou, and Zhicheng Dou. 2024. Grounding language model with chunking-free in-context retrieval. *arXiv preprint arXiv:2402.09760*.

Zile Qiao, Wei Ye, Tong Zhang, Tong Mo, Weiping Li, and Shikun Zhang. 2022. Exploiting hybrid semantics of relation paths for multi-hop question answering over knowledge graphs. In *COLING*, pages 1813–1822. International Committee on Computational Linguistics.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, et al. 2023. Webcpm: Interactive web search for chinese long-form question answering. *arXiv preprint arXiv:2305.06849*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *CoRR*, abs/2302.00083.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Ohad Rubin and Jonathan Berant. 2023. Long-range language modeling with self-retrieval. *CoRR*, abs/2306.13421.

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. *CoRR*, abs/2301.12652.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *NAACL-HLT*, pages 641–651. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi ..., and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. In *EMNLP*, pages 5657–5673. Association for Computational Linguistics.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024a. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.

Yujia Zhou, Zheng Liu, and Zhicheng Dou. 2024b. Boosting the potential of large language models with an intelligent information assistant. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. 2024c. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 1453–1463.

Yujia Zhou, Qiannan Zhu, Jiajie Jin, and Zhicheng Dou. 2024d. Cognitive personalized search integrating large language models with an efficient memory mechanism. In *Proceedings of the ACM on Web Conference 2024*, pages 1464–1473.