What is a Good Question? Assessing Question Quality via Meta-Fact Checking

Bo Zhang[1], Jianghua Zhu[1], Chaozhuo Li[2*], Hao Yu[1], Li Kong[1], Zhan Wang[1], Dezhuang Miao[3], Xiaoming Zhang[3], Junsheng Zhou[1]

[1]School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, China
[2]Key Laboratory of Trustworthy Distributed Computing and Service (MoE), Beijing University of Posts and Telecommunications, China
[3]School of Cyber Science and Technology, Beihang University, China

Abstract

Knowledge-based questions are typically employed to evaluate LLM's knowledge boundaries; meanwhile, numerous studies focus on question generation as a means to enhance the capabilities of both models and individuals. However, there is a lack of in-depth exploration about what constitutes a good question from the perspective of knowledge cognition. This paper proposes aligning the complete knowledge underlying questions with educational criteria effectively employed in physics courses, thereby developing novel knowledge-intensive metrics of question quality. To this end, we propose Meta-Fact Checking (MFC), which transforms questions into knowledge graph (KG) triples utilizing LLMs through few-shot prompting, thereby quantifying question quality based on the patterns observed within these triples. MFC introduces a novel interaction mechanism for KGs that communicates meta-facts, illustrating the types of knowledge that KGs can offer to the LLM for reasoning questions, rather than relying solely on the original triples. This strategy ensures that MFC remains unaffected by unexplored triples that LLM has not yet encountered within KGs compared to the retrieve-while-reasoning routine. Experiments across multiple datasets and LLMs demonstrate that MFC significantly improves the accuracy and efficiency of both question answering and assessing. This research marks a pioneering effort to automate the evaluation of question quality based on cognitive capabilities.

Introduction

Employing a variety of questions is a common approach to evaluate the knowledge boundary of large language models (LLMs), highlighting their efficacy in detecting and mitigating factuality hallucination (Ji et al. 2023; Huang et al. 2023). Furthermore, studies have generated questions through automated or semi-automated methods, serving as constructing benchmark datasets to specific domains (Gu et al. 2021a; Molina et al. 2024), educational scene (Kurdi et al. 2020; Wu et al. 2023) and model self-evaluation (Wang, Cho, and Lewis 2020; Zhang et al. 2023, 2024). It is essential to assess whether the utilized or generated questions possess the necessary quality for designated purposes.

The quest for effective methodologies in the automatic assessment of question quality continues to evolve. Current approaches combine metric-based assessments, including BLEU, ROUGE and METEOR, with human evaluations that consider fluency, relevance, and difficulty (Nema and Khapra 2018; Kurdi et al. 2020; Mulla and Gharpure 2023; Benedetto et al. 2023). They define high-quality questions mainly based on well-structured syntax and grammar, which is inadequate for effectively evaluating individuals or models' capabilities in knowledge-intensive tasks.

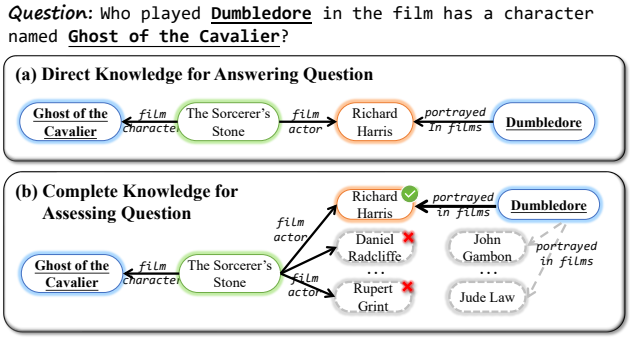


Figure 1: Knowledge requirement for question answering and question assessing.

In contrast to prior linguistic-based metrics, this paper advocates for designing knowledge-intensive metrics. To achieve this, we draw upon education-oriented criteria to define high-quality questions (Bates et al. 2014), which have proven effective in evaluating student-generated questions in physics courses. These criteria include *Cognitive Levels* that span from Recall to Analysis, the strategic inclusion of feasible *Distractors*, *Unambiguity* in the description of the question, and *Correctness* which involves a clear explanation of the correct solution. This evaluation framework takes into account individuals' *cognitive capabilities*, which emphasizes the relevant knowledge, how it is processed, and its retention in memory when responding to questions. Unfortunately, these criteria can only be assessed by domain experts, and varying interpretations may arise among different experts and disciplines. This variability presents challenges in quantitatively measuring question quality.

The principal aim of education-oriented criteria is to measure both knowledge density and knowledge correlation possessed by individuals. By developing knowledge-intensive

*Corresponding author: C. Li (lichaozhuo@bupt.edu.cn).

metrics that reflect the specific knowledge patterns underlying questions, these metrics can be effectively linked to individuals' cognitive capabilities. Knowledge graphs (KGs) standardize knowledge structuring by organizing factual information into triples, thereby allowing for a straightforward representation of knowledge as a set of triples. Therefore, a simple yet generic solution is to transform the questions into KG triples, where the patterns are analyzed to infer the question quality. A key aspect of the knowledge relevant to the questions is the direct triples required for formulating answers. For example, as shown in Figure 1(a), the question can be resolved using just three triples. However, the implicit knowledge is also essential in assessing overall question quality. The triples within the dashed boxes in Figure 1(b) are equally significant for reasoning about the connections between the entities `Ghost of the Cavalier` and `Dumbledore`. It requires to obtain the complete knowledge (i.e., direct and implicit) for assessing the question quality.

Recently, LLMs are extensively utilized in diverse reasoning tasks, with several studies utilizing LLMs to search the direct triples within KGs to facilitate question answering. One approach integrates Text-to-SQL generation with LLM reasoning to gather the direct triples (Xie et al. 2022; Jiang et al. 2023; Li et al. 2024). Another approach involves prompting the LLM to extend the reasoning path, starting from the question's topic entities and progressively identifying the entities that yield the answer (Sun et al. 2024; Guan et al. 2024; Jiang et al. 2024). Both approaches follow a routine: acquiring the reasoning path by sequentially selecting triples until the required knowledge is obtained. The LLM's reasoning for the subsequent triple is entirely contingent on the question and the already explored triples, which presents a challenge. It is problematic as the reasoning process is also influenced by unexplored triples (i.e., implicit knowledge) that the LLM has not yet encountered. For example, in Figure 1, the choice of which `(The Sorcerer's Stone, film actor, ?)` triples to explore further depends entirely on the LLM's inherent knowledge, despite the fact that KGs possess the relevant information. Harnessing complete knowledge from KGs enhances both question answering and question quality assessing.

To this end, this paper introduces a novel methodology to quantify the question quality towards cognitive capabilities. We design a set of knowledge-intensive metrics which are measured by comparing the patterns of complete knowledge triples against a predetermined set of reference patterns, offering a systematic framework for assessment. They provide a better grading to the questions across various cognitive capabilities. Furthermore, we introduce Meta-Fact Checking (MFC) method to address the significant challenge of acquiring triples that encompass complete knowledge by decoupling KG searching from LLM reasoning. This approach establishes a novel interaction between KGs and LLMs, where LLMs are provided with high-level meta-facts that embed the summary of the triples, rather than the triples themselves during searching phrase. LLMs then refine the questions based on the available meta-facts, iterating this process until an adequate collection of meta-facts is acquired. Finally, MFC makes informed decisions based on the complete knowledge available underlying the meta-facts. The main contributions are outlined as:

- This paper marks the first attempt to automatically evaluate question quality through the lens of cognitive capabilities, establishing knowledge-intensive metrics and introducing a benchmark dataset.
- We propose a plug-and-play method MFC to obtain complete knowledge from KGs, providing sufficient evidence for both question answering and question assessing.
- Experiments on multiple datasets with various LLMs confirm that MFC can significantly enhance LLM's ability to answer and assess questions, with the proposed metrics offering a more precise grading system compared to reasoning difficulty and relevance.

Metrics for Assessing Question Quality

Questions discussed in this paper only involve the knowledge-based questions which are formulated around domain-specific knowledge. Formally, a given question q is linked to a knowledge graph $\mathcal{G}$, composed of triples $\{(s, r, o)\}$, where s represents the head entity, r the relation, and o the tail entity. The goal is to convert the question q into the triples $Y_q \subset \mathcal{G}$ and then to assess the quality of q by analyzing the reasoning path patterns in Y_q. The triples Y_q involve the complete knowledge relevant to the question q, which constitutes an m-step reasoning path from the topic entities $\{s_0\}$ to the answer entities $\{o\}$ via relation chains $\{r_1, r_2, \cdots, r_m\}$. The intermediate entities are those within Y_q excluding $\{s_0\}$ and $\{o\}$. There may be one or multiple intermediate entities for a given entity e and the relation r.

Bates et al. (2014) enumerate multiple criteria to define high-quality questions around cognitive capabilities. We transform them into four knowledge-intensive metrics, as demonstrated below.

1. Cognitive Levels (COG) is a Bloom's Taxonomy metric (Bloom et al. 1956), which is a widely acceptable assessment model for the recall and processing of knowledge. It includes six cognitive levels: Recall, Understand, Apply, Analyze, Evaluate, and Create. This study excludes the "Evaluate" and "Create" levels due to the challenges in adequately supporting them with a single question. Table 1 illustrates the examples corresponding to four cognitive levels. The cognitive levels are aligned with distinct reasoning path patterns that divides questions into four categories as follows:

- **Recall**. The step m is 1. The triples Y_q only involve a relation r with the topic entity s_0 of the question q. They take the form (s_0, r, o) or (s_0, r^{-1}, o), where r^{-1} represents the inverse of r. All the possible entities $\{o\}$ are regarded as answers.
- **Understand**. Given an intermediate entity s_{i-1} and the subsequent relation r_i, the triples Y_q only involve a single entity s_i (excluding the answer entities $\{o\}$), i.e., (s_{i-1}, r_i, s_i) or (s_{i-1}, r_i^{-1}, s_i). The question requires understanding the relationships between entities to deduce the final answers.

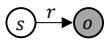
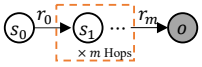
Levels	Examples	Reasoning Path Patterns
Recall	**Question 1**: Who influenced Samuel? **Triples**: (Samuel, *influenced_by*, Baruch Spinoza), (Samuel, *influenced_by*, Thomas Browne), $\cdots$ **Assessment**: It's a *Recall Question* as it can be answered directly by recalling a basic triple (Samuel, influenced_by, ?).	 $(s, r, \{o\})$
Understand	**Question 2**: Who was vice president under the subject of film Reagan? **Triples**: (Ronald Reagan, *film_subject.films*, Reagan), (Ronald Reagan, *vice_president*, George H. W. Bush) **Assessment**: It's an *Understand Question* as it requires comprehending the subject of the film "Reagan" before identifying his vice president.	 $(s_0, r_0, s_1), (s_1, r_1, s_2), \cdots$ $(s_m, r_m, \{o\})$
Apply	**Question 3**: Who played Dumbledore in the film has a character named Ghost of the Cavalier? **Triples**: (The Sorcerer's Stone, *film_character*, Ghost of the Cavalier), (The Sorcerer's Stone, *film_actor*, Daniel), (The Sorcerer's Stone, *film_actor*, Richard), $\cdots$, (Daniel, *film_character*, Harry Potter), (Richard, *film_character*, Dumbledore), $\cdots$ **Assessment**: It's a *Apply Question* as it requires necessitating an initial enumeration of all actors in The Sorcerer's Stone, followed by the application of knowledge about each actor to deduce their characters.	 $(s_0, r_0, s_1), (s_1, r_1, \{s_2\}), \cdots$ $(\{s_m\}, r_m, \{o\})$
Analyze	**Question 4**: How many people live in cities in the vicinity of the Nile? **Triples**: (Juba, *next_to_water*, Nile), (Kitchener's Island, *next_to_water*, Nile), $\cdots$, (Juba, *instance_of*, city), (Kitchener's Island, *instance_of*, river island) $\cdots$, (Juba, *population*, 459,342), $\cdots$ **Assessment**: It's an *Analyze Question* as it requires identifying all cities near the Nile and analyzing their population. The answer is the sum of identified city population.	 $(s_0, r_0, s_1), (s_1, r_1, \{s_2\}), \cdots$ $(\{s_m\}, r_m, \{o\})$ $\underline{analyze}(\{o\})$

Table 1: Examples of questions for different cognitive levels. Reasoning path patterns illustrate the general connection patterns of triples underlying the question. The dark node represents multiple entities $\{e\}$ while the light represents a single entity e.

- **Apply**. It is analogous to the Understand level except that s_i is not restricted to a single instance. This suggests that in addition to the comprehension required at the Understand level, resolving the question necessitates the application of knowledge concerning multiple entities $\{s_i\}$. This application is essential for progressing along the subsequent path by appropriately selecting the entities from $\{s_i\}$.
- **Analyze**. The answer entities in the set $\{o\}$ cannot directly address the question. In addition to the need for the Apply level, it is necessary to analyze the triples Y_q, which may involve operations such as counting, comparing, or contrasting the entities within $\{o\}$.

2. Unambiguity (UAM) is a binary metric assessing whether a question is clear or ambiguous. A low-quality question typically includes irrelevant or ambiguous information. For instance, in the question, "Among the countries where people speak Chinese, which one filmed the TV show The Bride with White Hair?", the information about speaking Chinese is redundant and ambiguous since The TV show was only filmed in China. Unambiguity involves questions that contains multiple topic entities. If any topic entity in $\{s_0\}$ is removed and the answer entities $\{o\}$ can still be fully deduced, the question is considered ambiguous.

3. Distractors (DTR) is a quantitative metric that indicates the number of distractors present in a question. The question quality also hinges on the presence of distractors. The user's intention in posing the question is to obtain the answer entity $\{o\}$. Distractors are defined as relations removed from the reasoning path that do not directly impact the determination of $\{o\}$. For example, in Question 2 of Table 1, which seeks the answer to "Who was the vice president of Ronald Reagan," the relation `film_subject.films` functions a distractor. Similarly, in Question 3, the distractor is "the film has a character named Ghost of the Cavalier." The number of distractors can be quantified by incrementally removing the relations in $\{r_{m-1}, \cdots, r_1\}$ until the answer entities $\{o\}$ are reached. The remaining relations serve to increase the difficulty of the question by adding distraction.

4. Correctness (CRT) is a binary metric, classifying the question as either correct or incorrect. The question must first ensure correctness, meaning the answer entities $\{o\}$ are non-empty. For example, the question "Among the countries where people speak Arabic, which one filmed the TV show The Bride with White Hair?" is incorrect.

Meta-Fact Checking

The objective of MFC is to obtain the triples Y_q from the knowledge graph that contains the complete knowledge of the question, which is essential for assessing the question quality. As shown in Figure 2, given the question q and the KG $\mathcal{G}$, MFC achieves the triples Y_q by searching meta-fact and refining reasoning path, followed by assessing metrics based on a predetermined set of reference patterns.

The process of MFC begins by prompting the LLM to automatically extract the topic entities $\{s_0\}$ to the question. During `Meta-Fact Search` stage, MFC reduces the original complex question q into a more straightforward sub-question sq_1, which revolves around a key entity s_0 from $\{s_0\}$. Subsequently, the KG $\mathcal{G}$ is queried to retrieve the candidate relations where s_0 serves as the head or tail

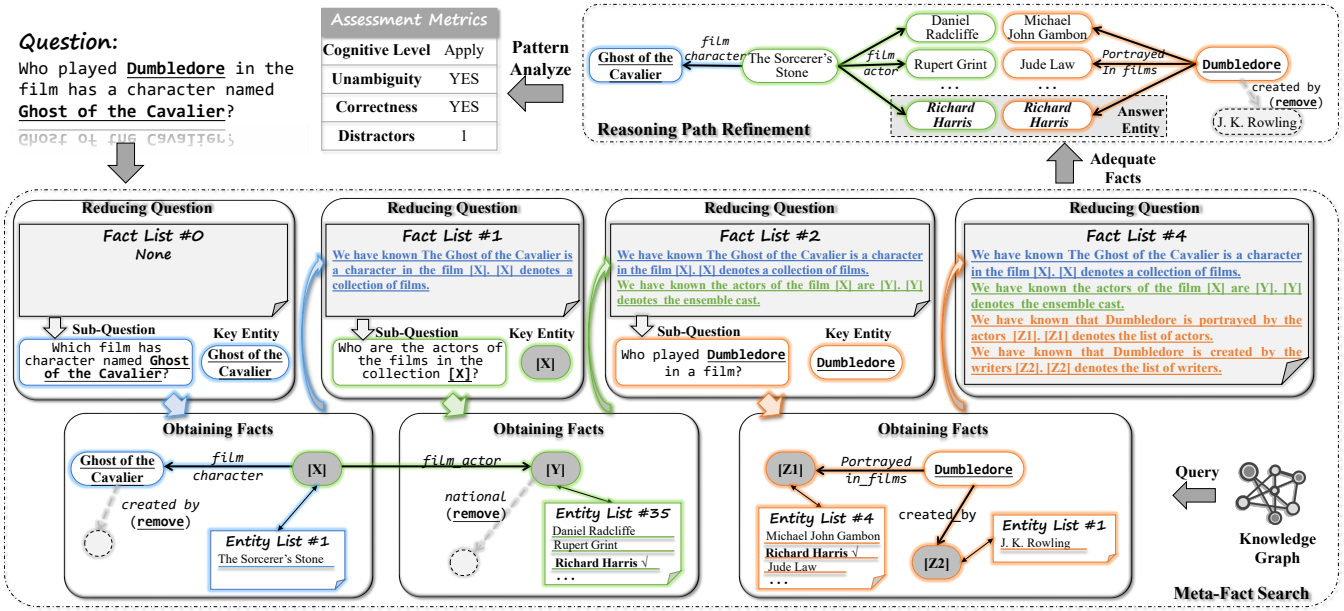


Figure 2: An overview of MFC for assessing question quality. It consists of two main components: Meta-Fact Search and Reasoning Path Refinement, which are driven by the LLM and KG under few-shot prompting.

entity. The LLM is then employed to select the relevant relations r_1 and generate the fact text that describes the contribution of the triples $(s_0, r_1, \{s_1\})$ or $(\{s_1\}, r_1, s_0)$ to sq_1. Here, $\{s_1\}$ denotes the intermediate entities linked by the same s_0 and r_1. The process iterates by producing the sub-question sq_2 based on the previously generated fact texts, with the key entity selected from $\{s_0\} \cup \{s_1\}$. This iterative process continues until sufficient fact texts are accumulated to address the original question q. Finally, the `Reasoning Path Refinement` stage unitizes the LLM to refine the reasoning path constructed from the obtained triples, involving the removal of irrelevant relations and the completion of missing ones.

Meta-Fact Search

One of main challenges for MFC is that the intermediate entities during reasoning are not unique. For example, the tail entities of `(The Sorceter's Stone, film_actor, ?)` encompass 35 entities, though only `Richard Harris` is relevant to the question. Identifying the key entity for the sub-question from numerous candidates proves to be impractical for the subsequent iterations, as it requires support from the additional triples not previously explored in KG. Prior work (Sun et al. 2024; Guan et al. 2024; Li et al. 2024) also suffers from this problem where they only depend on LLM's inherent knowledge, and thus KG's knowledge is not unitized sufficiently.

The main reason is that the question is formulated in natural language, which conveys the correlation between topic entities and answers through a sequence of relations (Atif, El Khatib, and Difallah 2023), e.g., (`film_character`$^{-1}$ $\rightarrow$ `film_actor` $\rightarrow$ `portrayed_in_film`$^{-1}$). The question itself lacks a natural connection with the intermediate entities. One of the practical solution is to hide the intermediate entities to enable the LLM to concentrate on searching triples based on the relation information, subsequently exposing all the intermediate entities to the LLM during reasoning process. Given this context, the intermediate entities generally exhibit similar object types, and are thus considered a meta-entity. Consequently, the LLM interacts with KGs by conveying meta-fact information that describes high-level triples containing the meta-entity, rather than the original triples. It contains two steps summarized below.

Reducing Question At the beginning of the D-th iteration, the inputs consist of the question q, the identified meta-facts F_D, and the current candidate entities E_D. The set E_D includes both meta-entities and topic entities that are not been explored in the previous iterations. The E_0 is initialized as $\{s_0\}$. Utilizing the identified meta-facts F_D, the LLM is prompted to evaluate whether F_D is adequate for solving the question q. If the evaluation yields a negative result, the LLM reduces the question q into a sub-question sq_{D+1} and identifies an associated key entity e_D within E_D. The sub-question sq_{D+1} is then forwarded to the `Obtaining Facts` step to gather knowledge relevant to sq_{D+1} by exploring e_D in the KG $\mathcal{G}$. Conversely, if a positive result is obtained, the process advances to the `Reasoning Path Refinement` stage.

To prevent infinite loops, the maximum number of iterations is set to D_{max}. If this limit is reached, it indicates the MFC's inability to explore more effective triples of q. The information obtained from $\mathcal{G}$ remains valuable for reasoning about the question q (Dai et al. 2024) and is subsequently utilized in the `Reasoning Path Refinement` stage.

Obtaining Facts Following the D-th iteration of `Reducing Question`, we are given the sub-question sq_{D+1} and its corresponding key entity e_D. The meta-facts F_{D+1} and the candidate entities E_{D+1} are initialized to F_D and E_D respectively. Initially, the KG $\mathcal{G}$ is queried to extract all candidate relations r that satisfy the conditions $(e_D, r, ?)$ and $(e_D, r^{-1}, ?)$, respectively. If e_D is a meta-entity, the relations r encompass those associated with the entities contained within the meta-entity. It is worth that the number of candidate relations for the meta-entity is comparable to that of a single entity due to the similarity in object types among the entities. The relations can be immediately obtained by executing two straightforward, pre-defined queries towards $\mathcal{G}$.

Subsequently, the LLM is utilized to pick out the top-K relevant relations r_{D+1} from the candidate set. For each relation r_{D+1}, the LLM is required to evaluate whether the triple (e_D, r_{D+1}, e') or $(e_D, {r_{D+1}}^{-1}, e')$ offers clear evidence relevant to the sub-question sq_{D+1}. If the evaluation is positive, the meta-fact text is generated by stating the fact information underlying the triples, which is included in the set F_{D+1}. Note here that a special token, "[X]," is utilized to denote the meta-entity that represents all the intermediate entities e', accompanied by a detailed explanation, when generating meta-facts. The candidate entities E_{D+1} are then updated by removing e_D and adding the new meta-entity [X]. If none of the r_{D+1} provide sufficient evidence, the meta-fact text stating "No relevant information is found for sq_{D+1}" is appended to F_{D+1}. It is important to note that, in cases where r_{D+1} and ${r_{D+1}}^{-1}$ are inverse relations, only one of these relations is retained. This determination is easily made by examining the union of the tail entities associated with each relation. Finally, the updated F_{D+1} and E_{D+1} are provided to the `Reducing Question` step for the next iteration.

Reasoning Path Refinement

The introduction of Reasoning Path Refinement stage is motivated by two primary factors: (1) The presence of irrelevant relations in the searched triples that should be removed (e.g., `created_by` in Figure 2). (2) The reasoning path might be incomplete, requiring the addition of new relations to improve its completeness. To address these factors, we utilize the intrinsic knowledge of the LLM to refine the reasoning path. Specifically, this involves reverting the meta-entities to their original entities and obtaining all the triples that are associated in `Meta-Fact Search` stage. The LLM is then prompted to infer the answer entities $\{o\}$ within the context of the question q and the obtained triples. Meanwhile, the LLM is required to elucidate the reasoning process by generating the reasoning path from topic entities to answer entities. If the LLM identifies new relations that are absent from the initial reasoning path, it is further utilized to complete the possible entities associated with these new relations. Ultimately, the quality of the question can be assessed automatically by contrasting the metric patterns with refined reasoning path. The whole inference process of MFC contains D searching stage as well as a refining stage, which needs at most $2D + 1$ calls to the LLM.

Experiments

We conduct an evaluation of MFC to verify its effectiveness in multi-hop reasoning and question quality assessment across diverse datasets and LLMs. The code and data are available at https://github.com/gregbuaa/qqa-mfc.

Experimental Setup

Multi-hop Reasoning Datasets and Baselines. To rigorously evaluate the multi-hop reasoning capabilities of MFC, we evaluate its performance on four extensively utilized Knowledge Base Question Answering (KBQA) datasets: *WebQSP* (Yih et al. 2016), *CWQ* (Talmor and Berant 2018), *Simple Question (SimQu)* (Bordes et al. 2015) and *GrailQA* (Gu et al. 2021b). Following the standard protocols (Atif, El Khatib, and Difallah 2023), we reported the results in terms of exact match accuracy (EM). Considering the costs associated with LLM API calls, we randomly selected approximately 200 questions from the validation set of each dataset for the evaluation process.

We compare with *standard prompting (IO)*, *Chain-of-Thought prompting (CoT)* and *Thought-of-Graph (ToG)*, each utilizing three in-context demonstrations. The IO directly prompts the LLMs to generate the answers for the given question. CoT (Wei et al. 2022) generates the answers by applying "step-by-step" reasoning chains. Both IO and CoT prompting rely solely on the inherent knowledge within the LLMs. Conversely, ToG prompts the LLMs to perform beam searches on KGs by iteratively exploring multiple possible reasoning paths on KGs (Sun et al. 2024).

Question Quality Dataset *CWQ-QQA* and Baselines. We have developed a question quality dataset named *CWQ-QQA*, derived from CWQ to evaluate the quality assessment capabilities of MFC. This dataset comprises 265 high-quality examples, each manually annotated with COG, UAM, DTR, and CRT labels. Results are reported in terms of micro-F1 scores. To establish a baseline for evaluating the accuracy of MFC concerning question quality, we fine-tuned a Llama2-7B model. The CWQ-QQA dataset was divided, with 40% allocated for fine-tuning the Llama2-7B model and the remaining 60% used for evaluation of both the Llama2-7B model and MFC.

LLM and KG Implementation. MFC can be directly applied to any LLMs that support few-shot prompting. In the experiments, we utilize the GPT-3.5-turbo and GPT-4-turbo models from OpenAI. To guarantee the reproducibility of the experiments, the temperature of the sampling is set to 0.2, and the maximum length of the generated text is set to 512. The parameter D_{max} is defined as $|E_0| * \texttt{Depth}$, where $|E_0|$ represents the number of topic entities. In all experiments, the default value for `Depth` is set to 3, taking into account that the interaction with KGs is proportional to $|E_0|$. The number of relevant relations `Top-K` is to set 3. Freebase (Bollacker et al. 2008) is used as the KG, comprising 0.9 billion triples following the exclusion of special tokens and non-English data (Lan and Jiang 2020).

	Methods	SimQu	WebQSP	GrailQA	CWQ	CWQ-QQA	Avg.
Finetuned SOTA	DiFaR^2 (Baek et al. 2023)	85.8*	65.3*	-	-	-	-
	Pangu (Gu, Deng, and Su 2023)	-	79.6*	75.4*	-	-	-
	DECAF (Yu et al. 2023)	-	82.1*	-	70.4*	-	-
GPT-3.5	IO Prompting	35.0	67.2	32.0	45.2	40.6	45.8
	CoT Prompting	37.5	67.2	35.2	47.0	45.4	46.5
	ToG (Sun et al. 2024)	51.7	72.2	66.2	51.8	43.2	57.2
	MFC (Ours)	**55.0**	**78.9**	**76.0**	**62.8**	**51.5**	**64.8**
	Gain over ToG	(+3.3)	(+6.7)	(+9.8)	(+11.0)	(+8.3)	(+7.6)
GPT-4	IO Prompting	38.0	68.7	41.6	47.8	45.3	48.3
	CoT Prompting	40.7	68.7	40.9	51.3	47.5	49.8
	ToG (Sun et al. 2024)	66.1	**78.6**	71.4	65.3	50.1	66.3
	MFC (Ours)	**67.3**	76.3	**76.5**	**72.2**	**63.1**	**71.1**
	Gain over ToG	(+1.2)	(-2.3)	(+5.1)	(+6.9)	(+13.0)	(+4.8)

Table 2: EM Accuracy on multi-hop reasoning. The fine-tuned state-of-the-art (SOTA) models provide references that are evaluated using the entire validation sets, rather than on the subsets.

Multi-hop Reasoning Results

Table 2 reports the EM accuracy of multi-hop reasoning applied to KBQA datasets. Although MFC does not require training, it achieves performance comparable to that of state-of-the-art (SOTA) methods that involve fine-tuning, despite being at a disadvantage in this comparison.

MFC consistently outperforms other prompting-based methods across various datasets, including SimQu, WebQSP, GrailQA, CWQ, and CWQ-QQA. MFC utilizes external KGs to enhance the knowledge base of LLM, resulting in significant improvements over CoT and IO methods, with average enhancements of 18.3% and 21.3% on GPT-3.5 and GPT-4, respectively. It indicates that the integration of KGs substantially benefits the reasoning capabilities of LLMs.

While MFC and ToG exhibit similar performance on SimQu and WebQSP, which involve one to two reasoning hops, MFC demonstrates clear advantages over ToG in the context of GrailQA, CWQ, and CWQ-QQA specifically designed for multi-hop reasoning. Furthermore, MFC requires less LLM call time compared to ToG (6.78 vs 11.15 avg. on all datasets). While ToG has a time complexity of $O(2ND + D + 1)$, MFC's complexity is $O(2D + 1)$, where D is the KG search depth and N is the beam search width. This improvement stems from the novel meta-fact information, which reduces the need for LLM reasoning over N entities at each depth level. These findings suggest that MFC is better at addressing multi-hop reasoning challenges, where the strategy of acquiring complete knowledge prior to deriving answers proves effective and efficient.

Question Quality Assessment Results

Table 3 reports the F1 score of question quality on the CWQ-QQA dataset we construct. The fine-tuned Llama2-7B demonstrates sub-optimal performance on the UAM and CRT classification, as these are closely linked to knowledge-based content. The metrics COG and DTR are evident in the linguistic syntax, as exemplified by the presence of statistical vocabulary in the Analyze questions, including terms such as "latest" and "how many." GPT-4 outperforms GPT-3.5 in question assessing for most metrics, benefiting from its enhanced reasoning capabilities.

Methods	COG	UAM	CRT	DTR
Llama2-7B	**73.2**	72.3	54.3	70.8
MFC (w/ GPT-3.5-turbo)	63.5	**86.2**	91.0	74.3
MFC (w/ GPT-4-turbo)	70.2	83.5	**96.4**	**76.8**

Table 3: F1 score of proposed metrics over CWQ-QQA.

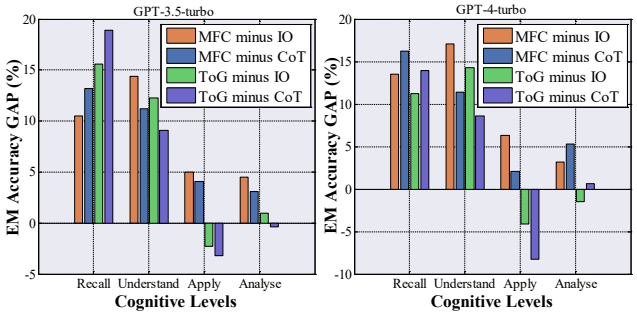


Figure 3: Average gap between KG-enhanced and LLM-driven methods with different cognitive levels.

Cognitive Levels Give Better Grading to Questions

Numerous studies utilize reasoning hops as a metric to evaluate the difficulty of questions (Atif, El Khatib, and Difallah 2023), akin to the DTR metric we introduced. However, it is often inadequate for many questions, as it only captures knowledge density while neglecting the intricacies of knowledge correlation. In this experiment, we investigate how different cognitive levels affect the grading of questions by employing ToG and MFC methods. Specifically, both ToG and MFC select the relevant triples from KGs and then leverage LLM to reason the answers by combining its inherent knowledge. The accuracy gap between the KG-enhanced models (ToG and MFC) and LLM-driven models (IO and CoT) highlights the challenges LLMs face in reasoning when utilizing KGs.

Motivated by these findings, Figure 3 shows the accuracy gap between KG-enhanced and LLM-driven methods across the categories of Recall, Understand, Apply and Analyze,

Methods	GPT-3.5-Turbo		GPT-4-Turbo	
	EM	Call	EM	Call
MFC (w/o meta)	54.7	9.5	68.8	9.3
MFC	62.8	8.7	72.2	8.4

Table 4: EM Accuracy and LLM call time of MFC and MFC (w/o meta) over CWQ.

as evaluated over the CWQ-QQA dataset. One can see that the gap diminishes as cognitive levels increase. The LLM's performance can be significantly enhanced by recalling one-hop triples from KGs. However, MFC yields only slight improvements, while ToG negatively affects the LLM's performance when addressing Apply and Analyze questions. Although the Understand, Apply, and Analyze categories involve similar reasoning hops, the difficulty of reasoning varies significantly among them. It concludes that cognitive levels give a better grading to multi-hop reasoning questions. In addition, our MFC has demonstrates superior accuracy compared to ToG when applied to Apply and Analyze questions. It provides the clear evidence that obtaining complete knowledge enhances the reasoning capabilities of LLMs.

Ablation Study

We carry out ablation studies to evaluate the significance of various settings in MFC.

Sensitivity of `Depth` and `Top-K`. To investigate the impact of expanding the search range on MFC's performance, we conduct an experiment to evaluate the accuracy of answer reasoning by varying the parameters `Depth` and `Top-K` within the set $\{1, 2, 3, 4, 5\}$ on the CWQ dataset and GPT-3.5-turbo. The corresponding average LLM call time is reported for each configuration. As shown in Figure 4, an increase in `Depth` results in improved performance for MFC; however, this enhancement is accompanied by an increase in computational costs. When the depth exceeds 4, the performance gains become marginal, and the frequency of LLM calls remains largely constant. It suggests that MFC does not engage in continuous interactions with KGs indiscriminately; instead, it halts the search once meta facts provide adequate evidence to the question. Conversely, MFC's performance declines significantly when the `Top-K` parameter exceeds 3. This reduction may be attributed to the presence of numerous irrelevant triples during searching stage, which disrupt MFC's decision-making process.

Effect of Meta-Fact Prompt Design. The meta-fact serves as a high-level abstraction that conveys information from KGs. To examine its impact, we ablate MFC as MFC (w/o meta) by removing the description of the meta-entity in the prompt, such as "[X] denotes a collection of films" in `Obtaining Fact` step. As reported in Table 4, the EM accuracy of MFC (w/o meta) shows a decline, while the LLM call time exhibits a significant increase. This finding suggests that detailed information for characterizing the extracted triples from KGs is crucial for guiding LLMs in refining its search direction in `Reducing Question` step.

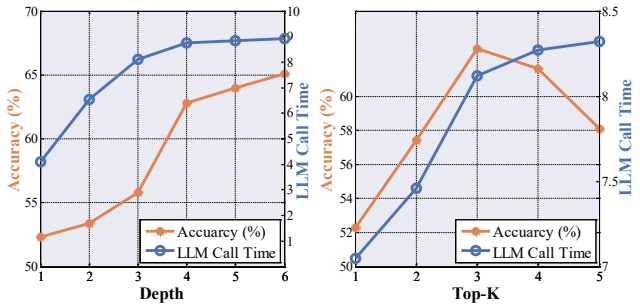


Figure 4: MFC's performance varying iterative `Depth` and `Top-K` relations over CWQ using GPT-3.5-turbo.

Related Work

Vanilla methods utilize established metrics for text generation, including BLEU, METEOR, and ROUGE, for the automated assessment of questions (Du and Cardie 2017; Nema and Khapra 2018). Human evaluation incorporates criteria such as fluency, relevance, naturalness, and difficulty, which are assessed through manual rating of the questions (Kurdi et al. 2020; Mulla and Gharpure 2023; Benedetto et al. 2023). However, the discussion surrounding the relationship between questions and knowledge is notably absent, which is essential for evaluating quality based on cognitive capabilities rather than superficial syntax and grammar.

LLMs offer the potential to achieve this by leveraging their advanced reasoning capabilities in conjunction with KGs. One area of research focuses on enabling LLMs to generate SQL queries for interacting with KGs during the reasoning process (Xie et al. 2022; Jiang et al. 2023; Li et al. 2024); however, this approach encounters challenges related to scalability and exhibits inflexibility in terms of knowledge updating. Another research direction involves utilizing LLMs to explore the neighboring entities associated with topic entities through both depth and breadth search strategies until adequate knowledge is acquired (Sun et al. 2024; Guan et al. 2024; Jiang et al. 2024). Nevertheless, these approaches are insufficient for acquiring the complete knowledge necessary to address the underlying questions, thereby limiting their applicability in evaluating question quality.

Conclusion

We propose novel evaluation metrics and corresponding solution MFC to answer the fundamental inquiry, "what is a good question?" in the context of integrating LLMs and KGs. The quality of a question is conceptualized as the cognitive capabilities that individuals should prioritize, necessitating the organization of complete knowledge to enhance reasoning regarding the questions. Experimental results suggest that MFC surpasses existing multi-hop reasoning methods and proves effective and efficient in quality assessment.

Limitations and Future Work. The metrics we propose are exclusively focused on knowledge-based questions, excluding items such as questionnaires and open knowledge questions. This framework will be further adapted for these types of questions in the future work.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 62406144, 62306145 and 62277031, and in part by Frontier Technologies R&D Program of Jiangsu (No. BF2024076).

References

Atif, F.; El Khatib, O.; and Difallah, D. 2023. Beamqa: Multi-hop knowledge graph question answering with sequence-to-sequence prediction and beam search. In *SIGIR*.

Baek, J.; Aji, A. F.; Lehmann, J.; and Hwang, S. J. 2023. Direct Fact Retrieval from Knowledge Graphs without Entity Linking. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *ACL*.

Bates, S. P.; Galloway, R. K.; Riise, J.; and Homer, D. 2014. Assessing the quality of a student-generated question repository. *Physical Review Special Topics-Physics Education Research*, 10(2): 020105.

Benedetto, L.; Cremonesi, P.; Caines, A.; Buttery, P.; Cappelli, A.; Giussani, A.; and Turrin, R. 2023. A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9): 1–37.

Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R.; et al. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 1247–1250.

Bordes, A.; Usunier, N.; Chopra, S.; and Weston, J. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Dai, X.; Hua, Y.; Wu, T.; Sheng, Y.; Ji, Q.; and Qi, G. 2024. Large Language Models Can Better Understand Knowledge Graphs Than We Thought. arXiv:2402.11541.

Du, X.; and Cardie, C. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2067–2073.

Gu, J.; Mirshekari, M.; Yu, Z.; and Sisto, A. 2021a. ChainCQG: Flow-Aware Conversational Question Generation. In *EACL*, 2061–2070.

Gu, Y.; Deng, X.; and Su, Y. 2023. Don't Generate, Discriminate: A Proposal for Grounding Language Models to Real-World Environments. In *ACL*, 4928–4949.

Gu, Y.; Kase, S.; Vanni, M.; Sadler, B.; Liang, P.; Yan, X.; and Su, Y. 2021b. Beyond IID: three levels of generalization for question answering on knowledge bases. In *WWW*.

Guan, X.; Liu, Y.; Lin, H.; Lu, Y.; He, B.; Han, X.; and Sun, L. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *AAAI*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, W. X.; and Wen, J.-R. 2023. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In *EMNLP*.

Jiang, J.; Zhou, K.; Zhao, W. X.; Song, Y.; Zhu, C.; Zhu, H.; and Wen, J.-R. 2024. Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. *arXiv preprint arXiv:2402.11163*.

Kurdi, G.; Leo, J.; Parsia, B.; Sattler, U.; and Al-Emari, S. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30: 121–204.

Lan, Y.; and Jiang, J. 2020. Query Graph Generation for Answering Multi-hop Complex Questions from Knowledge Bases. In *ACL*.

Li, X.; Zhao, R.; Chia, Y. K.; Ding, B.; Joty, S.; Poria, S.; and Bing, L. 2024. Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources. In *ICLR*.

Molina, I. L.; Švábenskỳ, V.; Minematsu, T.; Chen, L.; Okubo, F.; and Shimada, A. 2024. Comparison of Large Language Models for Generating Contextually Relevant Questions. *arXiv preprint arXiv:2407.20578*.

Mulla, N.; and Gharpure, P. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1): 1–32.

Nema, P.; and Khapra, M. M. 2018. Towards a Better Metric for Evaluating Question Generation Systems. In *EMNLP*.

Sun, J.; Xu, C.; Tang, L.; Wang, S.; Lin, C.; Gong, Y.; Ni, L.; Shum, H.-Y.; and Guo, J. 2024. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. In *ICLR*.

Talmor, A.; and Berant, J. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In Walker, M.; Ji, H.; and Stent, A., eds., *NAACL*.

Wang, A.; Cho, K.; and Lewis, M. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *ACL*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

Wu, Y.; Nouri, J.; Megyesi, B.; Henriksson, A.; Duneld, M.; and Li, X. 2023. Towards Data-Effective Educational Question Generation with Prompt-Based Learning. In *Science and Information Conference*, 161–174. Springer.

Xie, T.; Wu, C. H.; Shi, P.; Zhong, R.; Scholak, T.; Yasunaga, M.; Wu, C.-S.; Zhong, M.; Yin, P.; Wang, S. I.; Zhong, V.; Wang, B.; Li, C.; Boyle, C.; Ni, A.; Yao, Z.; Radev, D.;

Xiong, C.; Kong, L.; Zhang, R.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2022. UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models. In *EMNLP*.

Yih, W.-t.; Richardson, M.; Meek, C.; Chang, M.-W.; and Suh, J. 2016. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In *ACL*.

Yu, D.; Zhang, S.; Ng, P.; Zhu, H.; Li, A. H.; Wang, J.; Hu, Y.; Wang, W. Y.; Wang, Z.; and Xiang, B. 2023. DecAF: Joint Decoding of Answers and Logical Forms for Question Answering over Knowledge Bases. In *ICLR*.

Zhang, P.; Guo, J.; Li, C.; Xie, Y.; Kim, J. B.; Zhang, Y.; Xie, X.; Wang, H.; and Kim, S. 2023. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *WSDM*.

Zhang, X.; Peng, B.; Tian, Y.; Zhou, J.; Jin, L.; Song, L.; Mi, H.; and Meng, H. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *arXiv preprint arXiv:2402.09267*.