Counterfactual Question Generation Uncovering Learner Contradictions

Bo Zhang[1,2], Hao Yu[1], Wenjie Dong[1], Yvhang Yang[1], Dezhuang Miao[3], Fengyi Song[1,2], Yanhui Gu[1], Xiaoming Zhang[3], Junsheng Zhou[1,2*]

[1]School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, China
[2]Adolescent Education and Intelligence Support Lab of Nanjing Normal University, Laboratory of Philosophy and Social Sciences at Universities in Jiangsu Province, China
[3]School of Cyber Science and Technology, Beihang University, China
zhangbo@nnu.edu.cn, {f.song, gu, zhoujs}@njnu.edu.cn, {taiyue, yolixs}@buaa.edu.cn

Abstract

Conventional feedback, even when accompanied by brief explanations, rarely uncovers the hidden contradictions that trigger a learner's mistake. We bridge this gap with counterfactual question generation (CFQG): given a learner's answer, generate a follow-up question that deliberately contradicts it, compelling the learner to confront the underlying conflict. CFQG thus transforms assessment from passive scoring into an interactive and contradiction-centered dialogue that supports knowledge repair. To automate CFQG, we propose GapProbe, which probes the knowledge gap between a learner's belief and curated facts through a knowledge graph (KG), then designs counterfactual questions (CFQs) that negate the belief. Identifying contradiction-aware triples, and more importantly, selecting those most likely to confuse the learner, are highly challenging in large-scale KGs. GapProbe tackles these challenges with an iterative ProConB cycle coupled with a schema-aware KGMap. By caching one- and multi-hop schema patterns of the KG, KGMap provides "roadmap" to guide LLMs jump to deep and contradiction-aware triples, beyond traditional step-wise graph traversal. We present the CFQG benchmark and corresponding metrics for evaluating how generated CFQs trigger, focus, and deepen learner reflection through explicit contradictions. Experiments on multiple datasets and LLMs show that GapProbe boosts LLM reasoning over KGs and generates follow-up questions that consistently promote deeper and more focused learner reflection.

Introduction

Assessing learner answers is fundamental to judging knowledge mastery, memory consolidation, and higher-order reasoning. Yet most current assessment practices still rely on passive feedback such as binary judgments, numerical scores and brief explanations, which seldom encourage students to reflect on the reasoning behind their responses (Singh et al. 2018; Chang and Ginter 2024; Adlakha et al. 2024). To foster deeper cognitive engagement, a more constructive alternative is to recast learner answers as concise *counterfactual questions (CFQs)*, for example, *"If your (incorrect) answer were correct, what else would have to be true?"* These CFQs deliberately contradict established facts, exposing explicit *inconsistencies* in the learner's reasoning and prompting critical re-evaluation of initial beliefs. The approach extends the classic IRF (Initiation-Response-Follow-up) paradigm (Sinclair and Coulthard 1975) which shifts from explicit evaluation to guiding learners in testing their own hypotheses. By supporting interactive reflection, CFQ-based scaffolding helps learners repair faulty inferences, leading to deeper conceptual understanding.

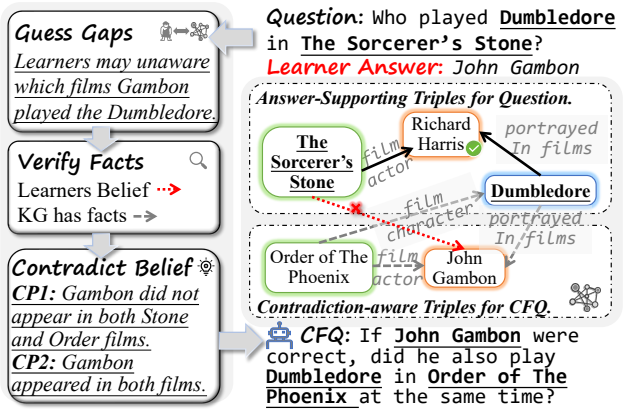

Figure 1: A KG-Supported CFQG Example.

Building on this pedagogical insight, we introduce **Counterfactual Question Generation (CFQG)** as a novel task that reconstructs the original question while integrating in the learner's answer to expose hidden contradictions. Presenting these contradiction-focused follow-ups compels learners to directly confront and resolve misaligned knowledge. For instance, when a learner claims "Gambon played Dumbledore in The Sorcerer's Stone," CFQG injects background knowledge (Figure 1) to highlight that only one actor portrayed the role in each film, thus challenging the erroneous assertion. While intuitively appealing, the formal definition of "inconsistency" and its automation in CFQG remain underexplored, which motivates this work.

Inconsistency in learner reasoning usually arises from *Contradictory Premises (CPs)*, i.e., cases where a learner's belief A (e.g., "Gambon portrayed Dumbledore in The Sorcerer's Stone") directly conflicts with the established fact $\neg A$ (e.g., "Harris was officially cast in that role"). Such

*Corresponding author

CPs inherently violate the Law of Non-contradiction (LNC), which forbids A and $\neg A$ from being simultaneously true. Knowledge graphs (KGs) transform this idea into a computable task: KGs supply the evidence needed to automatically verify logical consistency between learner beliefs and established knowledge. Leveraging this capability, our CFQG pipeline leverages KGs to detect CPs that violate LNC and then generates CFQs that explicitly surface these contradictions, encouraging learners to reflect on and correct mismatches in their understanding.

Recent advances in large language models (LLMs) have significantly improved knowledge-intensive reasoning in KG-based question answering (KGQA). Most existing methods primarily focus on retrieving answer-supporting facts from KGs by generating text-to-SQL queries (Xie et al. 2022; Jiang et al. 2023; Li et al. 2024) or extracting semantic paths (Sun et al. 2024; Miao et al. 2025; Zhang et al. 2025). While effective for standard KGQA, these approaches encounter two unique challenges when adapted to constructing CPs and generating CFQs. First, **identifying contradiction-aware triples**, i.e., those that expose knowledge gaps rather than simply support an answer, is substantially more difficult. As illustrated in Figure 1, the LLM must not only understand the context but also actively uncover latent contradictions. Second, **selecting confusion-focused evidence** remains challenging: even after contradictory facts are found, the model must prioritize those most likely to mislead the learner. Such confusion often arises when multiple plausible alternatives exist or when subtle distinctions between facts are critical. Consequently, transitioning from standard answer retrieval to contradiction-driven question generation demands deeper exploration into both probing knowledge gaps and constructing authentic contradictions.

We propose **GapProbe**, a novel framework designed for knowledge-grounded CFQG that unifies knowledge gap probing with efficient schema-aware KG retrieval. GapProbe hinges on an iterative **ProConB** cycle that (i) guesses the learner's potential knowledge gaps, (ii) retrieves relevant facts from KGs, and (iii) constructs contradiction-supporting premises for CFQG as illustrated in Figure 1. This alternation between focused guessing and evidence-based verification allows GapProbe to surface both answer-supporting and contradiction-aware triples through efficient trial-and-error search. The guessing phase of this cycle is especially challenging: modern KGs contain hundreds of millions of triples, making it difficult for an LLM both to align its guesses with real KG facts and to locate those facts most likely to mislead the learner. Therefore, we introduce a dynamic **KGMap** memory that steers the LLM toward high-priority schema regions and ranks candidate triples by their confusability. The main contributions are outlined as:

- This paper marks the first attempt to automatically generate follow-up CFQs that expose latent contradictions in learner responses, along with a new benchmark and reflection-oriented evaluation metrics for the CFQG task.
- We present a plug-and-play GapProbe that finds contradiction-aware evidence from KGs and adaptively targets knowledge most likely to confuse learners.
- We demonstrate through comprehensive experiments that GapProbe produces CFQs that promote deeper and more focused learner reflection.

Related Work

Question Generation (QG) and Evaluation. Language models are widely used to generate questions from textual passages (Du, Shao, and Cardie 2017) and knowledge graphs (Zhao et al. 2024), supporting benchmark construction for various domains (Gu et al. 2021a; Molina et al. 2024) and model self-evaluation (Wang, Cho, and Lewis 2020; Zhang et al. 2024). Quality is usually assessed with surface-form metrics such as BLEU, METEOR, and ROUGE (Du and Cardie 2017), supplemented by human ratings of fluency, relevance, and difficulty (Kurdi et al. 2020; Mulla and Gharpure 2023; Benedetto et al. 2023). More recent work introduces knowledge-intensive metrics that consider factual density and correlation (Zhang et al. 2025). Yet existing QG studies rarely address follow-up questions whose explicit goal is to surface contradictions in a learner's answer, and no established metric evaluates how effectively a question provokes such knowledge conflict.

LLM Reasoning with KGs. LLMs have become central to fact-intensive reasoning over KGs due to their strong language understanding and flexible query abilities. One common approach translates natural-language questions into executable SPARQL queries to retrieve triples directly (Xie et al. 2022; Jiang et al. 2023; Li et al. 2024), but scalability and continual KG updates remain problematic. Another research direction casts the LLM as an agent that decomposes a query and iteratively explores semantic paths through the graph (Sun et al. 2024; Guan et al. 2024; Tan et al. 2025; Miao et al. 2025). Here, the model incrementally extracts relevant neighborhood facts by traversing both depth and breadth, enabling it to assemble supporting evidence for questions. However, such approaches typically lack awareness of the underlying KG schema, causing the agent to perform blind or redundant exploration. In contrast, KGMap provides schema-level guidance, allowing for targeted retrieval avoiding training or blind exploration.

GapProbe Framework

This paper focuses on knowledge-grounded questions in domain-specific contexts, where both the original and counterfactual questions are structurally anchored to a KG $\mathcal{G} = \{(s, r, o)\}$ with entities $s, o \in \mathcal{E}$ and relations $r \in \mathcal{R}$. Given a learner's answer a_ℓ to a question Q, our goal is to generate a CFQ C that embeds the core proposition of a_ℓ and logically incorporates contradictory premises (CPs) derived from the KG. These CFQs are designed not only to reveal inconsistencies in the learner's reasoning but also to stimulate deeper reflection and self-correction.

To accomplish this, GapProbe employs an iterative **ProConB** cycle that probes and contradicts learner beliefs as shown in Figure 2. It consists of three key steps: (1) `Guess`: the LLM hypothesizes a piece of knowledge that the learner might be missing; (2) `Verify`: matching triples of guess step are retrieved from the KG; (3) `Contradict`: a CFQ

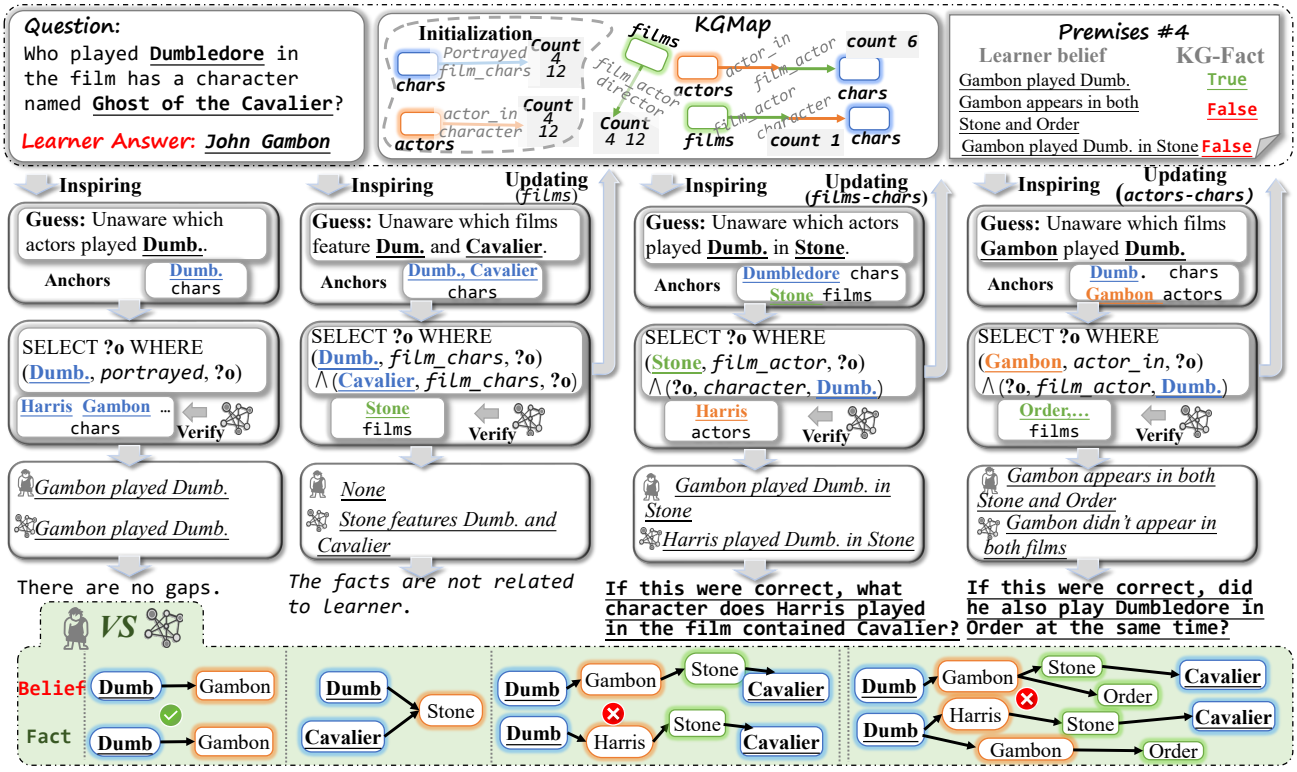

Figure 2: Illustration of the GapProbe Framework: Iterative ProConB Guided by KGMap Memory. The bottom dashed box compares the reasoning paths in the KG between the learner's possible belief and the supporting facts for each iteration.

is constructed to directly challenge the learner's belief by incorporating verified KG evidence to form a contradiction. This process is repeated, with each new iteration conditioned on accumulated premises, enabling GapProbe to generate increasingly challenging and insightful CFQs. Recognizing the inefficiency of blind guess search in large KGs, we introduce **KGMap**, a dynamic schema memory that guides ProConB toward high-impact, confusion-prone relations and entities. By combining schema-aware retrieval with iterative contradiction generation, GapProbe not only enhances the LLM's reasoning ability over KGs but also produces questions that foster deeper learner reflection.

KGMap: Profiling KG Schema

KGMap is designed to address two fundamental challenges in identifying knowledge gaps between a learner's answer and a large-scale KG. First, **LLMs typically lack structural awareness of the KG**, causing their guesses to drift toward irrelevant or even nonexistent facts. With hundreds of millions of triples, it is intrinsically difficult for the model to predict which facts would meaningfully contradict a learner's belief. Second, **not all contradictions are equally effective**. For instance, guessing "learners unaware which actors played in The Sorcerer's Stone" is overly broad, as the list is too extensive for meaningful reflection. In contrast, hypothesizing that "learners unaware which actors played Dumbledore" is much more precise, focusing the contradiction on a concise and memorable set of facts. Therefore, effective probing must target KG regions where contradictions are both plausible and maximally instructive.

To overcome these obstacles, KGMap serves as a dynamic schema memory that profiles both one-hop and multi-hop relation patterns in the KG. Before the initial ProConB iteration, KGMap identifies all topic entities $\{e_t\}$ from the original question Q and learner answer a_ℓ. For each topic entity e_t, KGMap queries $\mathcal{G}$ to obtain its type τ and enumerates all relations r associated with e_t, either as outgoing $(e_t, r, ?)$ or $(?, r, e_t)$. Entities of the same type usually share similar relational neighborhoods (for example, both `Harris` and `Gambon` are actors and thus have comparable relation patterns). For each type τ, the LLM selects the top-K relations most relevant to the question, storing these in a roadmap:

$$R = \{\tau \mapsto \{(r_1 : n_1), \cdots, (r_K : n_K)\}\} \quad (1)$$

where n_i denotes the cardinality (i.e., number of triples) under type τ that instantiate relation r_i. KGMap is updated dynamically at every `Verify` step. If a new entity type τ is encountered, its top-K relations are appended to R. When a retrieved path traverses a sequence of relations $r'_1 \cdots r'_h$ (e.g., `film_character` $\mapsto$ `film_actor`), the corresponding multi-hop pattern is also recorded as $\tau \mapsto (r'_1 \cdots r'_h : n'_h)$, where n'_h is the number of valid paths. This design allows subsequent ProConB iterations to use both one-hop and composite multi-hop relations as atomic query options.

By storing both one-hop and multi-hop schema patterns

and mapping entity types to their most relevant relations, KGMap enables the LLM to compose SPARQL queries that jump directly to deep, contextually relevant evidence, thus avoiding inefficient, blind, and incremental exploration prevalent in previous approaches (Sun et al. 2024; Zhang et al. 2025). Furthermore, KGMap prioritizes branches whose cardinality lies within an intermediate range (neither too small nor too large), which automatically steers probing toward regions rich in plausible alternatives and yields the most instructive counterfactual evidence. This schema-aware retrieval not only improves KG reasoning, but as shown in our experiments also enhances the model's self-correction and capacity for deeper reflection.

ProConB: Probing + Contradicting Beliefs

ProConB is an iterative framework guided by KGMap memory that systematically transforms a learner's answer into increasingly challenging and informative CFQs through a loop of guessing, verification, and contradiction, as follows:

Guess At iteration D, the LLM receives the current context $\mathcal{C}_D = \big(Q, a_\ell, \prod_{1:D-1}, R\big)$, where $\prod_{1:D-1}$ denotes the set of previously generated premises and R is the current KG roadmap. A task-specific prompt instructs the LLM to propose a hypothesis g_D, that is, a concise statement describing a possible gap in the learner's knowledge, along with its associated entity anchors $\{(\text{type}(e_D), e_D)\}$. Leveraging the top-$K$ relations and confusability statistics in R, the LLM is guided to generate high-relevance and KG-grounded guesses that are non-redundant with respect to prior steps.

Verify Given the hypothesis g_D, the LLM constructs a SPARQL query using the anchors $\{(\text{type}(e_D), e_D)\}$, the current entity table O_D and the relevant relations from R, where O_D maintains the IDs of all entities encountered up to the current iteration and is initialized with the topic entities, i.e., $O_0 = \{e_t\}$. Unlike prior methods limited to single-hop queries, our approach enables the LLM to compose multi-hop path queries by chaining up to h relations from $R[\tau]$, yielding expressions such as $r_1 r_2 \cdots r_h$. The resulting SPARQL query takes the form, for example, `SELECT` $?o$ `WHERE` $(e_o, r_1, ?x_1) \wedge \cdots \wedge (?x_{h-1}, r_h, ?o)$, where $e_o \in O_D$ and $r_1 \cdots r_h \in R$. This query retrieves end-point entities $\{e_D\}$ in a single call, thereby avoiding incremental expansions and reducing noise from intermediate entity nodes (e.g., $?x_1, \cdots, ?x_h$). To support confusability estimation, a `COUNT` clause is appended, producing $(r_1 \cdots r_h : n_D)$, where n_D records the number of discovered paths. If the query outputs $\{e_D\}$ introduce a new entity type τ', R is dynamically extended with $\tau' \mapsto (r'_1 : n'_1), \cdots, (r'_K : n'_K)$. Similarly, if a new multi-hop path is identified, the schema is augmented with $\tau \mapsto (r_1 \cdots r_h : n_D)$, enabling subsequent iterations to treat this composite relation as an atomic option. The entity table is also updated as $O_{D+1} = O_D \cup \{e_D\}$. This multi-hop querying capability streamlines KG access, allowing each guess to be resolved in a single SPARQL query and ensuring that the model efficiently retrieves precise, contextually relevant evidence for constructing contradictions. In practice, we set $h \leq 2$ to balance reasoning expressivity and computational efficiency.

Contradict When the `Verify` step yields novel triples, ProConB instantiates two premises. The *learner-belief premise* $\pi_L = (g_D, a_\ell)$ restates the learner's claim in the context of g_D. The *KG-fact premise* $\pi_K = (g_D, \{e_D\})$ pairs the same guess with the end-point entity $\{e_D\}$ retrieved from the KG. To ensure each follow-up is both focused and contradiction-driven, we filter premise pairs using two criteria: (1) π_L and π_K must be logically inconsistent (i.e., e_D must directly contradict a_ℓ), and (2) both premises must explicitly reference the learner's answer. Only premise pairs passing both checks are used to construct a CFQ, in which a_ℓ is embedded as a conditional clause and the contradiction is highlighted by querying about the critical element in π_K. The new premises are then appended to the running set, i.e., $\prod_{1:D} = \prod_{1:D-1} \cup \{\pi_L, \pi_K\}$.

This iterative process continues, updating the set of premises at each step, until either the maximum number of iterations $D_{\max}$ is reached or no further contradictory premises can be found. In this way, ProConB ensures that each generated CFQ remains tightly grounded in both the learner's prior answers and the evolving context, efficiently guiding the model toward deeper knowledge probing.

Evaluation and Benchmark Datasets

Reflection-Oriented Metrics: Trigger, Focus, Depth

A good CFQ *triggers* the learner's doubt with an explicit contradiction, *focuses* that doubt on knowledge tightly aligned to the original question, and *deepens* reflection by introducing just enough unfamiliar facts to require non-trivial reasoning. Let S_Q denote the answer-supporting triples for Q, S_C the triples for KG-fact premise π_K (i.e., answer-supporting plus contradiction-aware triples), and S_ℓ the explicit error triples for the learner answer (e.g., $\langle$*Stone, film_actor, Gambon*$\rangle$). For each sample in our dataset, both S_Q and S_ℓ are explicitly annotated to support autonomous evaluation. Following Zhang et al. (2025), we evaluate CFQs with three reflection-oriented metrics that quantify its impact on learner self-reflection.

Reflection Trigger (RT) evaluates whether the CFQ actively prompts the learner to reconsider their answer, requiring both conditions: *Premise Embedding* (E): The learner answer a_ℓ must appear verbatim or via a simple surface variant inside the CFQ stem. *Presence of Contradiction* (P): The CFQ's S_C contains at least one triple that shares a relation with a triple in S_ℓ, but differs in either the head or tail entity. A CFQ is considered to successfully trigger reflection if and only if both conditions hold: $RT = E \wedge P$.

Knowledge Focus (KF) quantifies how closely a CFQ's supporting facts align with the original question. An ideal CFQ should concentrate the learner's attention on facts directly relevant to resolving their initial misunderstanding, rather than diverging to unrelated information. For example, if the original question concerns casting, a CFQ focusing on award dates, though factually valid, would dilute the reflection value. Formally, for each triple $t \in S_C$, we compute its minimum semantic distance to any gold-standard triple in S_Q using a similarity function ψ (e.g., cosine similarity over text embeddings). The KF score is the fraction of triples in

S_C that fall within a threshold distance d_{thr} of S_Q:

$$KF = \frac{|\{t \in S_C | \min_{t' \in S_Q} (1 - \psi(t, t')) \leq d_{thr}\}|}{|S_C|} \quad (2)$$

A higher KF indicates that the CFQ maintains a tight fact link with the knowledge in the original question.

Cognitive Depth (CD) assesses the amount of new knowledge required by the CFQ. An effective CFQ should not only highlight inconsistencies but also encourage the learner to reason over additional facts or relations, thus deepening their cognitive engagement. For every triple $t = (s, r, o) \in S_C$ we assess whether it represents genuinely new information with respect to S_Q. A triple is considered "new" if it does not share both the relation and either the head or tail entity with any triple $t' \in S_Q$:

$$\text{amt}(t) = \begin{cases} 0, \exists t' : r = r' \wedge (s = s' \vee o = o'), \\ 1, \text{otherwise} \end{cases} \quad (3)$$

The CD score is the total amount of additional information it introduces, i.e., $\sum_{t \in S_C} \text{amt}(t)$. A higher CD indicates that the CFQ introduces more unfamiliar facts, thus requiring the learner to reason over a broader knowledge context.

Benchmark Dataset: *CWQ-CFQG*

We present benchmark dataset **CWQ-CFQG**, constructed atop CWQ benchmark (Talmor and Berant 2018) and grounded in real-world Freebase KG (Bollacker et al. 2008).

Dataset Overview. CWQ-CFQG contains 1,250 high-quality examples, each consisting of an original multi-hop question Q, a (possibly incorrect) learner answer a_ℓ, the set of answer-supporting KG triples S_Q, and a learner error set S_ℓ. Results are reported using our RT, KF, and CD metrics for autonomous evaluation, based on supporting triples S_C provided by models generating CFQs. Human evaluations are also introduced in our experiments.

Learner Answer Construction. To simulate realistic learner errors, we generate a_ℓ and S_ℓ using a stratified sampling procedure based on the number of plausible alternatives to the gold answer g. Specifically, for each example, we identify all triples $(s_i, r_i, g) \in S_Q$, and compute how many alternative candidate tails exist for each pattern $(s_i, r_i, ?)$ in Freebase. Triples are then grouped by the number of plausible alternatives: unique (only g is possible), moderate (several alternatives), and high (many possible alternatives). Triples are sampled from each group according to preset proportions, with an emphasis on moderate cases, and for each sampled triple, a plausible but incorrect answer is randomly chosen to construct a_ℓ and S_ℓ. This approach yields a diverse set of errors that reflect common learner confusions.

Train-Test Partitioning. Although GapProbe does not require training, CWQ-CFQG is partitioned to support both supervised and train-free evaluation in future research. **Training Set** includes 850 examples, each further annotated with 1–4 gold-standard CFQs spanning a range of CD scores, along with their supporting triples S_C. These CFQs are initialized by GapProbe and manually refined to ensure high quality. **Test Set** includes 400 examples reserved exclusively for evaluation.

	Methods	RT	KF	CD	Good (%)
GPT-4	MFC-CFQ	0.377	**0.967**	0.574	21.3
	w/o KGMap	0.505	0.937	0.892	42.3
	ours	**0.524**	0.963	**0.939**	**47.6**
DS-v3	MFC-CFQ	0.349	0.901	0.927	29.7
	w/o KGMap	**0.738**	**0.972**	0.748	49.5
	ours	0.672	0.960	**0.949**	**62.1**

Table 1: CFQ Quality Evaluation Results over CWQ-CFQG

CWQ-CFQG thus offers a challenging, well-annotated benchmark for evaluating the generation of CFQs that are reflection-triggering, knowledge-focused, and cognitively deep, all grounded in large-scale KGs and supporting both zero-shot and supervised evaluation paradigms. The data and code are available here https://github.com/gregbuaa/counterfactual-QG.

Experiments

Experimental Setup

CFQG baselines and evaluation. We evaluate on the test split of the proposed *CWQ-CFQG* dataset. Performance is assessed using both automatic metrics via **RT**, **KF** and **CD**, and human evaluation via A/B testing based on three aspects: Overall **OVR** (clarity and acceptability), Relevance **REL** (alignment with the original question), and Difficulty **DIF** (cognitive challenge). Baselines include: (1) *Llama2-7B fine-tuned*: a Llama2-7B model fine-tuned on CWQ-CFQG training set. (2) *MFC-CFQ*: an adaptation of MFC (Zhang et al. 2025) for CFQG, which applies the "ReduceQuestion-ObtainFacts" steps before generating CFQs. (3) *GapProbe w/o KGMap*: an ablation variant that replaces KGMap with the fact-gathering approach from MFC, omitting schema-driven SPARQL construction. When models generate multiple CFQs per example, we use the most recent iteration for comparison.

Multi-hop KG Reasoning. To assess the effectiveness of KGMap in KG reasoning, we also adapt GapProbe into a KBQA model, **GapProbe-QA**, by replacing the Guess and Contradict steps with a meta-fact summarization procedure similar to MFC. Experiments are conducted on four widely used KBQA benchmarks: *CWQ*, *SimpleQuestion* (Bordes et al. 2015), *GrailQA* (Gu et al. 2021b) and *CWQ-QQA* (Zhang et al. 2025). Following the standard protocols, we report **exact match (EM)** accuracy as the primary metric. Due to LLM API costs, we randomly sample 200 questions from each validation set for evaluation as in Guan et al. (2024). We compare GapProbe-QA with KG-free baselines *standard prompting (IO)* and *Chain-of-Thought prompting (CoT)* (Wei et al. 2022), as well as KG-based methods *ToG* (Sun et al. 2024) and *MFC*.

LLM and KG Implementation. GapProbe can be applied to any LLMs supporting few-shot prompting. We conduct experiments with Deepseek-v3 and GPT-4-turbo, using a temperature of 0.1 and a maximum generation length of 512 tokens for reproducibility. We set the similarity threshold d_{thr} in KF to 0.8. This threshold was selected through

Auto↓	DS-v3 OVR	DS-v3 REL	DS-v3 DIF	GPT-4 OVR	GPT-4 REL	GPT-4 DIF
RT	**0.448** **	-0.137	0.324 *	**0.522** ***	0.070	0.340 *
KF	-0.064	**0.672** ***	-0.194	0.124	**0.601** ***	0.018
CD	0.245	-0.046	**0.522** ***	0.362 *	0.157	**0.587** ***

Table 2: Kendall τ correlations between automatic metrics and human A/B testing between GapProbe and its w/o KGMap variant. n=100 pairs. *, ** and *** denote $p < \{.05, .01, .001\}$ respectively.

manual validation to ensure that only sufficiently relevant triples are considered, balancing specificity and relevance. The maximum iteration depth D_{max} is set to 5 while `Top-K` in KGMap is to set 3 for all experiments. Freebase is selected as the KG, containing 0.9 billion triples after filtering special tokens and non-English data (Lan and Jiang 2020).

CFQ Quality Evaluation Results

Table 1 reports the evaluation results of CFQG models on RT, KF, CD, and the `Good (%)` metrics, where `Good` is defined as the proportion of examples with RT$=1$, KF>0.30, and CD≥ 1. Both GapProbe and its w/o KGMap variant achieve substantially higher RT scores than MFC-CFQ, illustrating that our models are highly effective at generating CFQs that trigger learner reflection. Notably, the w/o KGMap variant attains KF scores nearly identical to those of GapProbe, showing that ProConB alone can maintain strong semantic alignment with the original question. However, GapProbe achieves markedly higher CD values, most prominently on DS-v3 (0.949 vs. 0.748), demonstrating its unique strength in producing CFQs with deeper cognitive challenge. This improvement is due to KGMap's schema-aware retrieval, which enriches the evidence and further directs the model to confusion-focused evidence, enhancing reasoning relevance and depth. Consequently, GapProbe delivers the highest proportion of "good" CFQs across both LLMs, highlighting its stability and overall effectiveness for reflection-oriented question generation.

Figure 3 summarizes our human A/B evaluation. We randomly sampled 100 examples from the test split and paired each GapProbe-generated CFQ with those from three baselines. Three independent annotators then judged each pair on OVR, REL and DIF, assigning a win or loss to GapProbe. To ensure reliability, we interspersed 20 control pairs and found over 70% agreement among the annotators. GapProbe overwhelmingly outperforms the fine-tuned baseline, highlighting the importance of KG integration, and still wins over 60% against MFC-CFQ and the w/o KGMap variant across all metrics. Table 2 further reports the Kendall τ correlations between automatic metrics and human A/B judgments. Each automatic metric aligns best with its target human assessment: RT with overall quality, KF with relevance, and CD with difficulty. This demonstrates that our automatic metrics

Models	CR (`cha`)	OCR (`cha`)↓	NI	SAA
Qwen2.5-14B				
w/o CFQ	19.0 (`38.0`)	**13.3** (`14.2`)	5.7	66.5
GapProbe	**30.4** (`70.9`)	20.8 (`35.0`)	**9.6**	**67.0**
Llama-3-70B				
w/o CFQ	8.70 (`21.7`)	**6.9** (`7.6`)	1.70	68.0
GapProbe	**33.3** (`62.3`)	8.4 (`11.5`)	**24.9**	**79.5**

Table 3: LLM-as-Student Simulation Results (%)

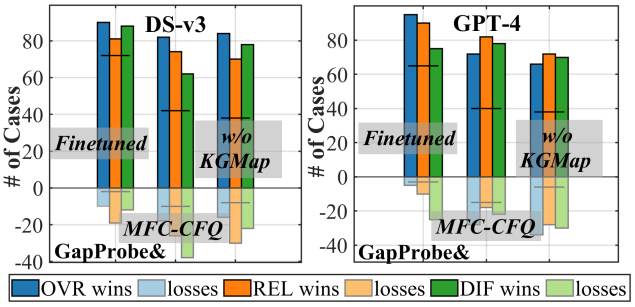

Figure 3: Human A/B results (N=100) comparing GapProbe vs. baselines on DS-v3 (left) and GPT-4 (right). Bars show win/loss counts for OVR, REL, and DIF; horizontal ticks mark cases where all three metrics favored GapProbe.

effectively capture the key aspects valued by human evaluators and are suitable for guiding reflection-oriented CFQG.

Simulated Learner Reflection with LLMs

To evaluate the impact of CFQs in uncovering learner contradictions, we simulate learner reflection using two models of varying ability: *Qwen2.5-14B* and *Llama-3-70B*, representing students at different proficiency levels. Each student LLM answers test questions, receives a model-specific CFQ for reflection, and may revise its answer based on self-assessment. As a baseline, the *No-CFQ* condition prompts the student LLMs to reconsider its answer without counterfactual feedback. We report six key metrics: *correction rate* (CR), the proportion of incorrect answers corrected after intervention; *over-correction rate* (OCR), the proportion of correct answers mistakenly changed; *CR cha* and *OCR cha*, the probability the model changes an answer (incorrect or correct, respectively) after intervention; *net improvement* (NI), CR minus OCR; and *self-assessment accuracy* (SAA), the proportion of cases where the model correctly judges its initial answer. Together, these metrics measure self-reflection effectiveness and meta-cognitive awareness.

Table 3 compares two LLMs under both No-CFQ and CFQ interventions. Notably, Llama-3 starts with a much lower baseline CR than Qwen2.5, reflecting greater overconfidence and reluctance to revise. However, with CFQ intervention, Llama-3's CR rises sharply to 33.3% (Δ+24.6pp), surpassing the improvement seen in Qwen2.5 (up to 30.4%, Δ+11.4pp). Although GapProbe leads to a modest increase in OCR, this can be attributed to the strong, often misleading, influence of CFQs. By encouraging deeper reflec-

	Methods	SimpleQuestion	GrailQA	CWQ	CWQ-QQA	Average
DS-v3	IO Prompting	32.5	26.8	45.0	40.4	36.2
	CoT Prompting (Wei et al. 2022)	32.5	30.0	49.4	45.3	39.3
	ToG (Sun et al. 2024)	**62.5**	68.1	56.3	51.7	59.7
	MFC (Zhang et al. 2025)	52.5	64.4	**60.0**	51.7	57.2
	GapProbe-QA (ours)	61.9	**75.6**	58.1	**57.0**	**63.2**
GPT-4	IO Prompting	33.8	26.3	42.5	41.9	36.1
	CoT Prompting (Wei et al. 2022)	33.1	25.6	48.1	44.5	37.8
	ToG (Sun et al. 2024)	62.5	70.6	55.0	**58.1**	61.6
	MFC (Zhang et al. 2025)	58.8	68.8	61.3	54.7	60.9
	GapProbe-QA (ours)	**68.8**	**78.8**	**64.4**	55.5	**66.9**

Table 4: Multi-hop KG reasoning Results over four benchmarks based on EM accuracy (%)

tion, CFQs prompt the model to reconsider even correct answers, leading to occasional over-correction. Nevertheless, the significant gains in CR far outweigh the slight rise in OCR, resulting in a strong overall net improvement. **These findings highlight the unique advantage of CFQ intervention: it effectively overcomes the overconfidence of larger LLMs, substantially improving their ability for self-correction and reflective learning.** Targeted CFQs deliver significant benefits, especially for stronger models less likely to reconsider their answers without such intervention.

Multi-hop KG Reasoning Results

Table 4 presents the EM accuracy of different methods for multi-hop reasoning across four KBQA benchmarks. KG-free prompting baselines (IO and CoT) perform considerably worse, emphasizing the necessity of explicit KG access for accurate multi-hop reasoning. Both MFC and ToG represent previous state-of-the-art KG-based approaches. GapProbe-QA leads to substantial gains over both baselines: on GPT-4, it increases from 60.9% (MFC) and 61.6% (ToG) to 66.9%. Improvements are especially notable on GrailQA and CWQ, the most challenging multi-hop datasets. Overall, these results highlight the effectiveness of KGMap's schema-aware and multi-hop retrieval in guiding the LLM to the most relevant graph regions, thereby enhancing both efficiency and accuracy. This demonstrates the broader utility of KGMap beyond CFQG, establishing it as an effective strategy for KBQA.

Ablation Study

Figure 4 presents the ablation results of GapProbe across DS-v3 and GPT-4, highlighting the impact of two key components: one/multi-hop SPARQL generation and whether or not confusability is considered (+P/-P) in Eq. (1). We evaluate four ablation baselines formed by combining these two components. The results of the four metrics are normalized to facilitate the visualization in the radar chart. It offers valuable insights into how each factor influences CFQ quality. Specifically, introducing confusability filtering substantially improves RT and Good metrics. This is because confusability directs the model's focus toward contradictions that are more likely to prompt reflection. However,this improvement comes at a slight cost in KF, as the broader contradictions generated can somewhat shift the focus away from the most relevant knowledge. slightly deviating from

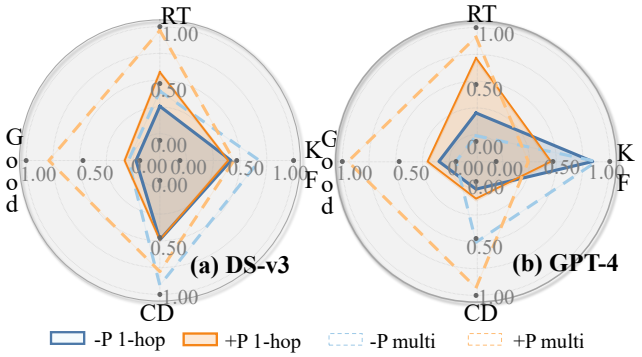

Figure 4: Ablation Study for GapProbe by ablating the components of multi-hop SPARQL generation and confusability.

the most relevant knowledge. On the other hand, multi-hop reasoning significantly enhances CD by allowing the model to access more complex, multi-hop knowledge, which leads to CFQs that require deeper reasoning. As a result, GapProbe with multi-hop reasoning and confusability (+P multi) generates a higher proportion of "Good" CFQs. Overall, these findings demonstrate that combining confusability filtering with multi-hop reasoning greatly enhances CFQ quality, making GapProbe more effective in promoting learner self-correction and fostering deeper reflective learning.

Conclusion

In this work, we proposed GapProbe, a novel framework for CFQG that integrates LLMs with KGs through an iterative, schema-aware reasoning process. Extensive experiments demonstrate three main advantages of our approach. First, GapProbe significantly enhances LLM reasoning ability over KGs. Second, GapProbe generates CFQs that actively promote deeper reflection and critical thinking in learners, going beyond surface-level factual recall. Third, results from our LLM-as-Student simulation further show that CFQs can effectively prompt model-based students to identify and revise their own misconceptions, leading to improved self-reflection and answer correction. These findings collectively highlight the potential of counterfactual prompts to advance both the reasoning capability and meta-cognitive awareness of AI systems in educational contexts.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (No. 62406144, 62277031, 22033002, 92370127), in part by Frontier Technologies R&D Program of Jiangsu (No. BF2024076), and in part by Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 24KJB520017).

References

Adlakha, V.; BehnamGhader, P.; Lu, X. H.; Meade, N.; and Reddy, S. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12: 681–699.

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.

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*.

Chang, L.-H.; and Ginter, F. 2024. Automatic short answer grading for Finnish with ChatGPT. In *AAAI*.

Du, X.; and Cardie, C. 2017. Identifying where to focus in reading comprehension for neural question generation. In *EMNLP*.

Du, X.; Shao, J.; and Cardie, C. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. In *ACL*.

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

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*.

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*.

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*.

Miao, D.; Du, Y.; Li, X.; Zhang, X.; Li, J.; Zhang, B.; Yan, B.; Zhang, L.; and Zhang, L. 2025. Reverse Chain-of-Thought and Causal Path Verification: A Modular Plugin for Aligning LLMs with Knowledge Graphs. In *CIKM*.

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.

Sinclair, J.; and Coulthard, M. 1975. *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford, UK: Oxford University Press.

Singh, P.; Sheorain, S.; Tomar, S.; Sharma, S.; and Bansode, N. 2018. Descriptive answer evaluation. *International Research Journal of Engineering and Technology (IRJET)*, 5(05): 2395–0056.

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 *NAACL*.

Tan, X.; Wang, X.; Liu, Q.; Xu, X.; Yuan, X.; and Zhang, W. 2025. Paths-over-graph: Knowledge graph empowered large language model reasoning. In *WWW*.

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*.

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*.

Zhang, B.; Zhu, J.; Li, C.; Yu, H.; Kong, L.; Wang, Z.; Miao, D.; Zhang, X.; and Zhou, J. 2025. What is a Good Question? Assessing Question Quality via Meta-Fact Checking. In *AAAI*.

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*.

Zhao, R.; Tang, J.; Zeng, W.; Chen, Z.; and Zhao, X. 2024. Zero-shot Knowledge Graph Question Generation via Multi-agent LLMs and Small Models Synthesis. In *CIKM*.