A Feedback-Driven Self-Evolving Knowledge Graph Framework for Automatic Question Generation

Bingzheng Yan
Beijing Zhongguan Academy
s-ybz25@bjzgca.edu.cn

Xin Zou
Beijing Zhongguan Academy
zouxin@bjzgca.edu.cn

Xin Liu
Beijing Zhongguan Academy
v-lx@zgci.ac.cn

Jian Li
Beijing Zhongguan Academy
lijian@bjzgca.edu.cn

Abstract

Automatic Question Generation (AQG) is valuable for educational assessment and everyday learning applications. However, most existing approaches rely on static knowledge extraction and single-pass LLM-based generation, a paradigm that often introduces substantial randomness and results in unstable question quality and limited consistency. To address this issue, we propose a feedback-driven **self-evolving knowledge graph** question generation framework (**SEKG-QG**), which continuously improves both knowledge graph quality and question generation performance through an evaluable and iterative closed loop. Given commonly used instructional documents (e.g., PDFs) as input, the framework first extracts entities and relations from the text and constructs an initial knowledge graph with an LLM. It then uses a stronger LLM to build a high-confidence reference knowledge graph as an evaluation benchmark, and quantitatively evaluates the initial graph from multiple dimensions such as entity and relation coverage and accuracy. Next, questions are generated based on the current knowledge graph and evaluated by the stronger LLM from several perspectives, including knowledge coverage, semantic consistency, and entity–relation correctness, while also producing structured revision feedback for the questions. Finally, this feedback is traced back to update the initial knowledge graph, yielding a corrected graph and entering a new round of graph evaluation, question generation, and quality evaluation, completing one full self-evolution iteration of the **SEKG-QG** framework. Experimental results show that this feedback-driven mechanism can significantly improve both knowledge graph quality and question generation quality, and demonstrate the stability and robustness of the framework across multiple AQG test settings. Overall, **SEKG-QG** organically integrates large language models with knowledge graphs, providing a general framework for high-quality and stable **AQG**.

1 Introduction

Automatic Question Generation (AQG) has been widely used in educational assessment and everyday learning applications, aiming to automatically transform instructional documents into a set of questions for evaluating learning outcomes [Heilman and Smith, 2010, Kurdi et al., 2020, Bikaun et al., 2024]. By automatically generating high-quality questions, AQG systems can reduce the cost of manual question writing at scale and support continuous learning assessment. Unlike tasks such as summarization or dialogue generation, AQG requires higher correctness and stability: once questions contain factual errors or omit key information, they may not only reduce the system's credibility but also mislead learners and negatively impact teaching outcomes [Kurdi et al., 2020, Zhou et al., 2023].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

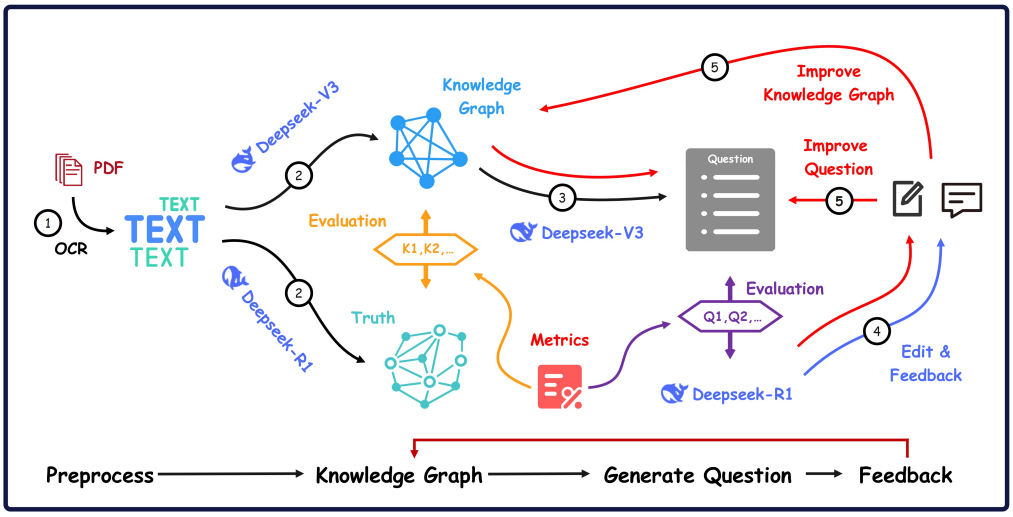

Figure 1: Technical roadmap of the feedback-driven **SEKG-QG** framework. The system extracts text paragraphs from PDF documents and constructs an initial knowledge graph with the **DeepSeek-V3** model. It further uses a stronger model (**DeepSeek-R1**) to construct a high-confidence reference knowledge graph and evaluate the initial graph with quantitative metrics. Then, it generates initial questions from the initial knowledge graph and evaluates them with DeepSeek-R1, producing both metric-based quality scores and structured revisions/feedback for the initial questions. Finally, these revisions and feedback are traced back to correct the knowledge graph, forming an iterative closed loop so that the knowledge graph and the generated questions improve together (Note: the numbers indicate the execution order. The iterative loop corresponds to the combination of the red arrows and the purple bidirectional evaluation arrows in Fig. 1. The order is step 3, the purple bidirectional evaluation arrows, step 4, step 5, and then the process returns to step 3 to start a new iteration).

In recent years, large language models have significantly improved the fluency and diversity of automatically generated questions [Raffel et al., 2020, Lewis et al., 2020, Brown et al., 2020]. However, in practice we observe that question quality produced by existing AQG systems is often unstable: performance can vary across subjects, and even under the same input conditions, repeated generations may yield large fluctuations in quality. Many methods adopt a static pipeline of "extract knowledge first, then generate questions": the system extracts entities and relations from the original text to build a fixed knowledge representation, and then uses an LLM to generate questions based on that representation. Once the early-stage knowledge extraction is biased, the errors may be amplified in subsequent generation, while existing systems lack effective mechanisms to correct them. Similar error accumulation has been widely reported in knowledge graph construction and knowledge-driven generation tasks [Pan et al., 2017, Zhang et al., 2018, Pan et al., 2020], and further impacts the knowledge consistency and reliability of generated content [Wang et al., 2022].

A key reason for these issues is that existing AQG systems typically lack an explicit, iterative quality evaluation and feedback mechanism. Without an effective feedback loop, a system cannot reliably determine whether the current questions cover key information or remain consistent with the underlying knowledge representation, causing knowledge-level errors to accumulate across later generations. Although recent research explores LLM-based self-reflection or self-improvement strategies [Bai et al., 2022, Madaan et al., 2023, Shinn et al., 2023], these methods mainly focus on revising the generated text itself and do not explicitly update the underlying structured knowledge representation. As a result, they struggle to fundamentally mitigate knowledge inconsistency, especially for AQG tasks that are highly sensitive to knowledge structure.

To this end, we propose a feedback-driven self-evolving knowledge graph framework for question generation (**SEKG-QG**), whose overall technical route is shown in Fig. 1. The core idea is to build an evaluable and iterative closed-loop process so that knowledge representations and question generation quality can mutually reinforce and co-evolve. Specifically, given commonly used instructional

documents (e.g., PDFs) as input, the system first extracts entities and relations from the text and constructs an initial knowledge graph with an LLM. It then introduces a stronger LLM to build a high-confidence reference knowledge graph and quantitatively evaluates the initial graph at both entity and relation levels. On this basis, the system generates questions from the current knowledge graph and performs multi-dimensional quality evaluation on the generated results, including knowledge coverage, semantic consistency, and correctness of entities and relations, while outputting structured feedback. This feedback is not only used to revise the generated questions but is also explicitly traced back to the knowledge graph layer to update and correct the original knowledge representation, forming an iterative loop of "knowledge evaluation–question generation–quality feedback–knowledge correction". By repeatedly executing this process, **SEKG-QG** can gradually improve the knowledge graph quality and generate more stable and higher-quality questions. Experiments on various instructional documents show that the framework improves both knowledge representation quality and question quality while reducing generation randomness, validating its stability and generality for AQG.

Our main contributions are as follows:

- We propose a feedback-driven self-evolving knowledge graph AQG framework (**SEKG-QG**) that enables continuous correction and evolution of knowledge representations via a closed-loop feedback mechanism;
- We design a two-level quality evaluation and structured feedback mechanism for both knowledge graphs and question generation, so that improvements in question quality explicitly drive knowledge structure updates;
- Experiments on multiple instructional documents show that the proposed framework substantially outperforms existing baselines in knowledge representation quality, question generation quality, and generation stability.

Our code is open-sourced at: `https://github.com/undoubtable/KG_allprocess`.

2 Related Work

2.1 Automatic Question Generation and Limitations of Knowledge Modeling

AQG is an important research direction in natural language processing, widely used in educational assessment and everyday learning applications. With the rapid development of large language models (LLMs), AQG has increasingly been modeled as a sequence-to-sequence generation task and trained end-to-end on large-scale data, significantly improving the fluency and diversity of generated questions [Raffel et al., 2020, Lewis et al., 2020, Brown et al., 2020].

However, existing AQG methods still show clear limitations in knowledge modeling and consistency guarantees. Most methods adopt a single-pass paradigm and rely on static knowledge extraction and representation. Once the underlying knowledge representation contains noise or omissions, the generation model often lacks effective mechanisms to detect and correct them, producing factually inconsistent or unreliable questions in knowledge-intensive settings [Pan et al., 2020, Wang et al., 2022]. This indicates that relying solely on end-to-end generation models is insufficient to meet the requirements of AQG for knowledge accuracy and generation stability.

2.2 Knowledge Graph-based Question Generation

To mitigate the limitations of end-to-end generation models in knowledge modeling, prior work has introduced structured knowledge representations such as knowledge graphs into AQG. Knowledge graphs explicitly model knowledge using entities and relations, providing structured and interpretable constraints for downstream question generation. Prior studies show that knowledge graphs can improve question relevance and factual correctness to some extent [Zhang et al., 2018, Chen et al., 2021].

However, most KG-based AQG methods assume the knowledge graph is static and reliable once constructed. In practice, knowledge graphs are often obtained via automatic extraction or weak supervision and inevitably contain noise, omissions, or structural biases. Due to the lack of mechanisms to trace back and correct the knowledge graph using generation outcomes, errors at the knowledge

representation level may be directly inherited or even amplified during question generation. Therefore, introducing structured knowledge alone does not fundamentally solve the problem; effective AQG systems still require mechanisms to continuously refine knowledge representations.

2.3 Feedback-Driven Learning and Knowledge Refinement

Feedback-driven learning has been widely studied in machine learning, including self-training, reinforcement learning, and reinforcement learning from human feedback (RLHF). Prior work shows that introducing feedback signals can effectively improve generation models in alignment, reliability, and overall usability [Ouyang et al., 2022, Bai et al., 2022].

Meanwhile, human-in-the-loop methods have also been applied to knowledge graph construction, refinement, and verification, where human feedback is leveraged to improve the quality of knowledge acquisition and representation [Bikaun et al., 2024]. Although these approaches emphasize the importance of feedback for knowledge quality control, they often heavily rely on explicit human supervision, which is costly and difficult to scale to large or multi-domain settings. Moreover, existing work typically separates feedback mechanisms from downstream generation tasks and lacks a systematic way to directly use feedback produced during generation to drive iterative refinement of explicit knowledge representations.

2.4 Knowledge Graphs and Large Language Models

With the rapid progress of LLMs, an increasing number of studies explore combining knowledge graphs (KGs) with LLMs to enhance reasoning ability, factual consistency, and controllable generation [Yasunaga et al., 2021, Yao et al., 2023, Sun et al., 2023]. In these studies, KGs are often used as external memory or structured guidance, providing more grounded and interpretable support for generation [Liu et al., 2024, Wang et al., 2024]. These results suggest that integrating symbolic knowledge with neural language models has strong potential for knowledge-intensive tasks.

However, most existing methods still treat KGs as fixed inputs and do not explicitly model the dynamic evolution of knowledge representations driven by generation outcomes. Meanwhile, the strong semantic understanding and evaluation ability of LLMs has seldom been leveraged as structured feedback signals to update underlying KGs. Different from prior work, we focus on building a feedback-driven closed-loop framework: LLMs serve not only as question generators and quality evaluators but also as sources of structured feedback that drives continuous KG evolution, thereby jointly improving knowledge representation quality and AQG performance.

3 Problem Definition

We study a feedback-driven self-evolving knowledge graph question generation framework (Fig. 1), whose core goal is to continuously improve the quality of the KG during iterations using structured feedback produced by LLMs, and to generate high-quality question sets.

Given a set of commonly used instructional documents (e.g., PDFs):

$$\mathcal{D} = \{d_1, d_2, \ldots, d_N\}, \tag{1}$$

for each document d_i, the text is first extracted via Optical Character Recognition (OCR), and then partitioned page by page into a set of text paragraphs:

$$\mathcal{P} = \{p_1, p_2, \ldots, p_M\}. \tag{2}$$

Based on the text, an initial knowledge graph $G^{(1)} = (E^{(1)}, R^{(1)})$ is constructed using a chosen base LLM, where $E^{(1)}$ denotes the extracted entity set and $R^{(1)}$ denotes the relation set between entities. The KG at iteration t is denoted as $G^{(t)} = (E^{(t)}, R^{(t)})$.

To quantify the quality of knowledge representations, we assume that a stronger LLM can be used to construct a high-confidence reference knowledge graph (a **benchmark**) $G^* = (E^*, R^*)$, and a set of KG quality metrics $\mathcal{M}_K^{(t)}$ (e.g., entity coverage, relation coverage, and entity/relation correctness) is used to evaluate $G^{(t)}$.

Then, based on $G^{(t)}$ and the original text paragraphs, the base LLM generates a question set $Q^{(t)}$. These questions are subsequently evaluated by a stronger LLM (question quality metrics $\mathcal{M}_Q^{(t)}$), which also produces corresponding structured revision feedback signals $\mathcal{F}^{(t)}$.

The core problem we focus on is how to effectively use the feedback signal $\mathcal{F}^{(t)}$ to iteratively update $G^{(t)}$, i.e.,

$$G^{(t)} \rightarrow G^{(t+1)}, \tag{3}$$

so that knowledge representation quality and AQG performance can be continuously improved together across iterations.

4 The SEKG-QG Framework

We propose **SEKG-QG**, a framework for jointly optimizing knowledge representations and question generation performance for AQG. The framework adopts a feedback-driven iterative closed-loop pipeline, including KG construction, KG-based question generation, LLM-based quality evaluation, and feedback-driven KG self-evolution. Through multiple iterations, the underlying knowledge representations are continuously corrected and evolved, thereby gradually improving the stability and overall quality of AQG.

Different from prior methods that treat the KG as a static input, **SEKG-QG** explicitly uses structured feedback produced during downstream question evaluation to continuously correct and improve the KG, forming a feedback-centric closed-loop evolution system.

4.1 Knowledge Graph Construction

Given the set of text paragraphs extracted from documents, we use the **DeepSeek-V3** model for entity recognition and relation extraction to construct an initial knowledge graph:

$$G^{(1)} = (E^{(1)}, R^{(1)}), \tag{4}$$

where $E^{(1)}$ denotes the extracted entity set and $R^{(1)}$ denotes the extracted relation set.

At this stage, the construction process prioritizes *coverage* rather than precision, and thus intentionally allows a certain amount of noise. This noise is expected to be gradually corrected in later iterations through feedback-driven knowledge refinement.

To enable reliable evaluation, we further adopt a stronger LLM and a more conservative extraction prompt to construct a high-confidence reference knowledge graph (benchmark):

$$G^* = (E^*, R^*). \tag{5}$$

We emphasize that G^* is not human-annotated ground truth, but serves as a benchmark for KG quality evaluation.

4.2 Knowledge Graph Evaluation Metrics

At iteration t, the current KG is:

$$G^{(t)} = (E^{(t)}, R^{(t)}). \tag{6}$$

To quantitatively evaluate KG quality, we define a set of knowledge-level evaluation metrics $\mathcal{M}_K$ under a *relaxed matching* setting [Paulheim, 2017, Hogan et al., 2021]. This setting allows us to ignore lexical variations to some extent, better fitting LLM-based knowledge extraction scenarios and reasonably capturing semantic correspondences between entities and relations.

Entity-related metrics Entity precision and recall characterize correctness and coverage completeness of entity extraction. Precision reflects extraction quality by penalizing redundant or incorrect entities, while recall reflects knowledge coverage by penalizing missed entities.

- **Entity Precision**

$$\mathrm{Prec}_E = \frac{|E^{(t)} \cap E^*|}{|E^{(t)}|}. \tag{7}$$

- **Entity Recall**

$$\mathrm{Rec}_E = \frac{|E^{(t)} \cap E^*|}{|E^*|}. \tag{8}$$

  Under relaxed matching, entity recall can also be viewed as *entity coverage*.

- **Entity F1**

$$\mathrm{F1}_E = \frac{2 \cdot \mathrm{Prec}_E \cdot \mathrm{Rec}_E}{\mathrm{Prec}_E + \mathrm{Rec}_E}. \tag{9}$$

Relation-related metrics

- **Relation Precision**

$$\mathrm{Prec}_R = \frac{|R^{(t)} \cap R^*|}{|R^{(t)}|}. \tag{10}$$

- **Relation Recall**

$$\mathrm{Rec}_R = \frac{|R^{(t)} \cap R^*|}{|R^*|}. \tag{11}$$

  Under relaxed matching, relation recall can also be viewed as *relation coverage*.

- **Relation F1**

$$\mathrm{F1}_R = \frac{2 \cdot \mathrm{Prec}_R \cdot \mathrm{Rec}_R}{\mathrm{Prec}_R + \mathrm{Rec}_R}. \tag{12}$$

Summary Entity-level and relation-level metrics jointly characterize the correctness of entities and relations in the KG. In experiments, we report each metric for fine-grained analysis; we can also aggregate them into a scalar KG quality score $K^{(t)}$ to characterize overall improvement trends across iterations [Paulheim, 2017].

4.3 Knowledge Graph-based Question Generation

In each iteration, the system generates a set of questions based on the current KG $G^{(t)}$ and the original text paragraphs:

$$Q^{(t)} = \{q_1^{(t)}, q_2^{(t)}, \dots\}. \tag{13}$$

The goal of this stage is to maximize the coverage of entities and relations in the KG while maintaining consistency between questions and the knowledge structure. Therefore, we use a moderately sized LLM, **DeepSeek-V3**, to balance generation quality and computational efficiency, rather than relying on the strongest available model.

4.4 Question Quality Evaluation Metrics

The generated question set $Q^{(t)}$ is automatically evaluated by a stronger LLM, **DeepSeek-R1**, which serves as an evaluator [Liu et al., 2023, Zheng et al., 2023b]. Given the current KG $G^{(t)} = (E^{(t)}, R^{(t)})$, the evaluator assesses each question $q \in Q^{(t)}$ from multiple complementary dimensions.

Question-level quality metrics We define four question quality metrics normalized to $[0, 1]$ as follows ($\mathcal{E}_q$ and $\mathcal{R}_q$ denote the sets of entities and relations appearing in the stem and options, respectively):

- **Knowledge Coverage (A)** Measures the proportion of entities and relations involved in a single question that are covered by the current KG, reflecting how much the question utilizes the underlying knowledge representation:

$$A(q) = \frac{\left|\left(\mathcal{E}_q \cap E^{(t)}\right) \cup \left(\mathcal{R}_q \cap R^{(t)}\right)\right|}{\left|E^{(t)} \cup R^{(t)}\right|}. \tag{14}$$

- **Semantic Coherence (B)** Measures semantic and logical consistency between the stem and options, scored by the evaluator based on clarity and the plausibility of options [Liu et al., 2023]:

$$B(q) \in [0, 1]. \tag{15}$$

- **Entity Alignment Accuracy (C)** Measures whether entities appearing in the question can be correctly aligned to entities in the current KG, reflecting entity-level consistency between the question and knowledge representation. Let $\mathcal{E}_q^{\text{align}} \subseteq \mathcal{E}_q$ denote the subset of entities in question q that the evaluator judges can be correctly aligned to some entity in $E^{(t)}$:

$$C(q) = \frac{|\mathcal{E}_q^{\text{align}}|}{|\mathcal{E}_q| + \epsilon}. \tag{16}$$

  where $\epsilon > 0$ is a sufficiently small constant for numerical stability.

- **Relation Correctness (D)** Measures whether the relations implied by the question are consistent with relations defined in the KG, reflecting factual correctness at the relation level. Let $\mathcal{R}_q^{\text{corr}} \subseteq \mathcal{R}_q$ denote the subset of relations implied by the correct answer that the evaluator judges to be semantically consistent with relations in $R^{(t)}$:

$$D(q) = \frac{|\mathcal{R}_q^{\text{corr}}|}{|\mathcal{R}_q| + \epsilon}. \tag{17}$$

The above metrics are aggregated into an overall question quality score:

$$Q(q) = w_A A(q) + w_B B(q) + w_C C(q) + w_D D(q), \tag{18}$$

where w_A, w_B, w_C, w_D are non-negative weights. We use uniform weights by default: $w_A = w_B = w_C = w_D = 0.25$. While outputting quantitative metrics $\mathcal{M}_Q^{(t)}$, the evaluator also generates structured textual feedback:

$$\mathcal{F}^{(t)} = \{f_1^{(t)}, f_2^{(t)}, \dots\}, \tag{19}$$

which points out issues such as missing entities/relations or ambiguous formulations, and guides KG updates in the next iteration.

Suite-level evaluation Beyond evaluating individual questions, we further assess the entire question suite $Q^{(t)}$ from a global perspective to measure whether coverage distributions over entities and relations are reasonable. In AQG, if questions concentrate on a small subset of entities or relations, this may introduce coverage bias and reduce the representativeness and effectiveness of the question set for teaching and assessment [Kurdi et al., 2020].

Based on this, we compute empirical distributions $P_{\mathcal{E}}$ and $P_{\mathcal{R}}$. We further define target distributions $P_{\mathcal{E}}^*$ and $P_{\mathcal{R}}^*$ to describe an ideal coverage distribution (e.g., uniform, or a prior induced by the KG structure).

The structural quality of the entire suite is defined as the match between empirical and target distributions:

$$S(Q^{(t)}) = 1 - \frac{1}{2}\left(\text{Dist}(P_{\mathcal{E}}, P_{\mathcal{E}}^*) + \text{Dist}(P_{\mathcal{R}}, P_{\mathcal{R}}^*)\right), \tag{20}$$

where $\text{Dist}(\cdot)$ denotes a symmetric distribution divergence. We adopt the Jensen–Shannon divergence to ensure boundedness and stability [Lin, 1991]. A higher $S(Q^{(t)})$ indicates that the coverage distribution of the question suite over entities and relations is closer to the target distribution, resulting in a more balanced structure and more sufficient coverage.

4.5 Feedback-Driven Knowledge Graph Self-Evolution

In each iteration, the feedback signal $\mathcal{F}^{(t)}$ produced during evaluation is parsed into structured knowledge refinement operations, including entity correction, relation correction, and missing-knowledge completion. These operations are applied to the current KG to obtain an updated graph:

$$G^{(t+1)} = \Phi\big(G^{(t)}, \mathcal{F}^{(t)}\big), \tag{21}$$

where $\Phi(\cdot)$ denotes the feedback-driven knowledge update operator. The updated KG $G^{(t+1)}$ enters the next iteration, continuing to generate questions with **DeepSeek-V3** and to be evaluated by **DeepSeek-R1**. Through multiple iterations, the KG gradually evolves toward higher coverage and higher correctness, leading to continuous improvements in AQG performance.

5 Experiments

5.1 Experimental Setup

We evaluate the proposed **SEKG-QG** framework on a set of educational assessment PDFs (40 PDF documents split from a law textbook). Each PDF is first parsed into text paragraphs, which serve as inputs for KG construction and question generation.

For each document, we compare two states before and after applying **SEKG-QG**: the initial KG $G^{(1)}$ and its generated question set $Q^{(1)}$, and the refined KG $G^{(2)}$ and its corresponding question set $Q^{(2)}$ after one round of feedback-driven self-evolution. Except for whether feedback-driven refinement is enabled, the remaining experimental settings are kept identical.

5.2 Evaluation Metrics

We evaluate the method from two levels: knowledge representation and question generation.

Knowledge graph metrics KG quality is evaluated using precision, recall, and F1 at both entity and relation levels, using the high-confidence reference KG constructed by DeepSeek-R1 as the benchmark. Metric definitions are provided in Section 4.2, and the aggregated score is denoted as $K^{(t)}$.

Question generation metrics Question generation quality is evaluated by DeepSeek-R1 as an automatic evaluator, with metric definitions in Section 4.4. The final question quality score $Q^{(t)}$ is computed by Eq. 18, and the suite-level score $S(Q^{(t)})$ is further computed by Eq. 20 to measure the coverage balance over entities and relations.

5.3 Knowledge Graph Quality Comparison

We first analyze how the SEKG-QG framework affects KG quality before and after refinement. We compute KG quality statistics across 40 PDFs; to highlight the contrast, we only report the results after one iteration (see Table 1). We observe consistent improvements at both entity and relation levels: the refined KG quality $\mathcal{M}_K^{(2)}$ is consistently better than the initial KG quality $\mathcal{M}_K^{(1)}$.

In particular, entity recall (coverage) increases from 0.792 to 0.858, indicating that the refinement process effectively completes missing entities in the initial KG. The improvement at the relation level is more substantial: relation F1 increases from 0.171 to 0.320, showing that refinement can significantly improve the accuracy and completeness of relation structures.

Table 1: Comparison of KG quality metrics $\mathcal{M}_K^{(t)}$ before and after applying SEKG-QG

Method	Prec_E	Rec_E	$\mathrm{F1}_E$	Prec_R	Rec_R	$\mathrm{F1}_R$
$\mathcal{M}_K^{(1)}$	0.724	0.792	0.778	0.197	0.151	0.171
$\mathcal{M}_K^{(2)}$	**0.768**	**0.858**	**0.810**	**0.429**	**0.304**	**0.320**

5.4 Generated Question Quality Comparison

We again use the same 40 PDFs to evaluate the impact of KG self-evolution on question quality. The question generation setting is to generate up to 50 questions per PDF; in practice, we obtain 40 suites totaling 1,672 questions. We average the results and report them in Table 2. We can see that although the improvement in knowledge coverage $A(q)$ is small, questions generated from the refined KG show clear improvements in semantic coherence, entity alignment accuracy, and relation correctness. On the aggregated metric, the average question quality score $Q(q)$ increases from 75.33 to 77.13, and the average suite-level score $S(Q^{(t)})$ increases from 76.42 to 79.29, indicating that the generated suites become more reliable and better balanced in coverage.

Table 2: Comparison of question quality metrics $\mathcal{M}_Q^{(t)}$ before and after applying SEKG-QG

Method	A(q)	B(q)	C(q)	D(q)	$Q(q)$	$S(Q^{(t)})$
$\mathcal{M}_Q^{(1)}$	85.64	83.25	63.28	69.14	75.33	76.42
$\mathcal{M}_Q^{(2)}$	**86.84**	**84.33**	**66.67**	**70.67**	**77.13**	**79.29**

5.5 Case Study

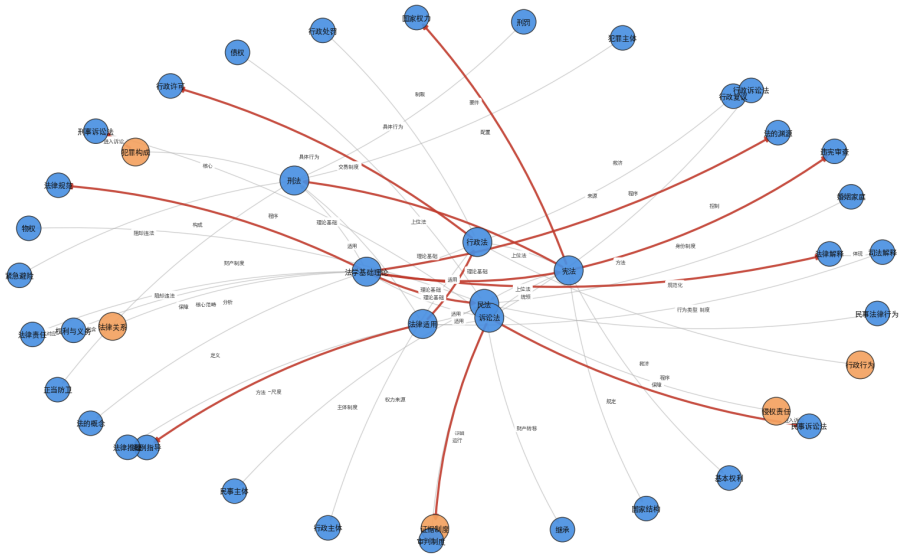

Figure 2: A randomly selected comparison of a knowledge graph before and after evolution. Blue nodes and gray edges indicate entities and relations in the initial KG; orange nodes indicate newly added nodes after KG self-evolution; red edges indicate corrected relations. This example shows that SEKG-QG can improve KG quality, consistent with Table 1: fewer entity edits and more relation corrections.

Fig. 2 presents a randomly selected document-level comparison of the KG and generated questions before and after applying SEKG-QG. We observe that missing or incorrect relations in the initial KG are corrected, and the corresponding generated questions become more accurate in entity references and relation statements, achieving better consistency with the underlying knowledge structure.

6 Discussion

The proposed SEKG-QG framework suggests that feedback-driven self-evolving knowledge graphs for question generation constitute an effective paradigm that can improve both structured knowledge representations and AQG quality. Unlike traditional pipeline methods that treat the KG as a static intermediate artifact, we model knowledge representations as dynamic objects that can be evaluated, corrected, and continuously evolved based on generation outcomes.

Why feedback-driven evolution works Our results reveal a key phenomenon: errors exposed during question generation often directly reflect structural defects in the underlying KG, such as missing entities, incorrect relations, or inappropriate abstraction granularity. By introducing an LLM

as an evaluator, SEKG-QG can transform issues observed during generation into actionable structured feedback and use it to drive iterative KG evolution [Paulheim, 2017, Hogan et al., 2021]. As a result, the KG and question generation process form a closed loop, enabling the system to correct structural knowledge rather than only making local adjustments at the generation strategy or decoding level.

The trade-off between coverage and precision Our experiments also show an inherent trade-off between knowledge coverage and structural precision. The initial KG is typically coverage-oriented and inevitably introduces noise, especially at the relation level. The feedback-driven self-evolution process can gradually reduce this noise and improve relation precision and overall consistency without significantly sacrificing coverage. This suggests that, given a reliable feedback mechanism, allowing a certain amount of noise in the early stage to gain higher coverage can be a reasonable and effective strategy.

The role of LLMs as evaluators Compared with traditional automatic metrics or rule-based verification, LLMs provide more flexible and semantically rich evaluation capabilities, enabling detection of fine-grained issues such as entity mismatch and logical inconsistency [Liu et al., 2023]. In **SEKG-QG**, LLMs serve not only as generators but also as evaluators and feedback providers, approximating human review at a relatively low marginal cost while producing structured feedback that can be directly used for knowledge updates. This observation suggests substantial potential for using LLMs as feedback sources in knowledge-centric systems.

Limitations Despite stable improvements, **SEKG-QG** has several limitations. First, relying on stronger LLMs for evaluation and reference KG construction introduces additional computational cost, and the evaluator itself may introduce model biases [Zheng et al., 2023a]. Second, the reference KG G^* is not guaranteed to be absolute ground truth, which may affect the interpretability of some quantitative results. Third, the current feedback parsing and refinement operator $\Phi(\cdot)$ is relatively coarse-grained; future work could incorporate confidence modeling and uncertainty estimation to improve the granularity of knowledge updates.

Future directions Future work can extend the framework in several directions:

- Designing adaptive stopping criteria, e.g., automatically determining the number of evolution iterations based on the convergence of KG quality or question quality improvements;
- Reducing the cost of LLM-based evaluation and feedback generation via evaluator distillation or lightweight models;
- More tightly coupling KG structure optimization with the question generation objective, e.g., explicitly incorporating knowledge coverage and question quality metrics into a unified optimization procedure;
- Generalizing the feedback-driven closed-loop evolution paradigm to other knowledge-intensive generation tasks, such as explanation generation, curriculum organization, and educational content recommendation.

Overall, the framework can be extended to other knowledge-dependent generation tasks such as explanation generation and educational content recommendation. SEKG-QG provides a general mechanism for improving the reliability and controllability of LLM-centered systems with structured knowledge.

7 Conclusion

We propose **SEKG-QG**, a feedback-driven self-evolving knowledge graph framework for automatic question generation. The framework builds an iterative closed loop of "knowledge graph construction, knowledge-guided question generation, LLM-based automatic evaluation, feedback-driven KG self-evolution, and re-generation", enabling continuous and joint improvements in KG quality and AQG performance.

Experiments show that applying SEKG-QG significantly improves entity coverage, relation correctness, and overall structural quality, and further yields more reliable and semantically consistent questions, including improvements in semantic coherence, entity alignment accuracy, and relation

consistency. Unlike existing methods that rely on static knowledge representations or single-pass generation, **SEKG-QG** provides a feedback-loop mechanism that traces back knowledge defects exposed during generation to iteratively refine explicit knowledge representations.

More broadly, our study suggests that treating structured knowledge as a dynamic component that can be evolved and corrected, rather than a fixed intermediate artifact, can substantially enhance the robustness and controllability of knowledge-intensive generation systems. We believe the proposed feedback-driven paradigm provides a solid and extensible foundation for deeply integrating LLMs with structured knowledge across broader applications.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Tyler Bikaun, Michael Stewart, Wei Liu, et al. Cleangraph: Human-in-the-loop knowledge graph refinement and completion. *arXiv preprint arXiv:2405.03932*, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Yifan Chen, Yifan Wu, Rui Zhang, et al. Knowledge-aware question generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Michael Heilman and Noah A. Smith. Good question! statistical ranking for question generation. In *Proceedings of NAACL*, 2010.

Aidan Hogan et al. Knowledge graphs. *ACM Computing Surveys*, 2021.

Ghada Kurdi et al. A systematic review of automatic question generation for educational purposes. *Education and Information Technologies*, 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 1991.

Qiang Liu, Shuohuan Wang, Shikun Feng, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. Knowgpt: Knowledge graph-based prompting for large language models. In *Advances in Neural Information Processing Systems*, 2024.

Yang Liu, Dan Iter, Yichong Xu, et al. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023.

Aman Madaan et al. Self-refine: Iterative refinement with large language models. In *Proceedings of ACL*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Shirui Pan et al. Knowledge graph construction techniques. *IEEE Transactions on Knowledge and Data Engineering*, 2017.

Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, and Heng Ji. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 2017.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

Noah Shinn et al. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Zequn Sun, Yikang Shen, Jiaxin Chen, Qian Wang, and Jiawei Han. Knowledge-augmented large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Xinyu Wang, Yue Zhang, Zhiyuan Liu, et al. Fact-guided neural text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Zhen Wang, Yuanning Cui, Zequn Sun, and Wei Hu. A prompt-based knowledge graph foundation model for universal in-context reasoning. In *Advances in Neural Information Processing Systems*, 2024.

Yuan Yao, Yuning Mao, and Jiebo Luo. Kglm: Integrating knowledge graphs into large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Yejin Choi. Qa-gnn: Reasoning with language models and knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Zhen Zhang, Nan Yang, Furu Wei, and Ming Zhou. Knowledge graph-based question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023a.

Lianmin Zheng et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023b.

Wenxuan Zhou et al. Large language models for educational question generation: Opportunities and challenges. *arXiv preprint arXiv:2306.03882*, 2023.