一种基于反馈驱动的知识图谱自进化题目生成框架

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

自动出题（Automatic Question Generation, AQG）在教学测评与日常教育应用中具有重要价值。然而，现有方法大多依赖静态知识抽取和单一大模型出题，这种范式往往引入较大的生成随机性，导致题目质量不稳定且一致性不足。为此，本文提出一种反馈驱动的知识图谱自进化题目生成框架（**SEKG-QG**），通过可评估、可迭代的闭环机制，结合反馈信号持续提升知识图谱质量与题目生成效果。以常用教学文档（如PDF）为输入，该框架首先抽取文本中的实体与关系并结合大语言模型构建初始知识图谱，随后，利用更强的大语言模型构建高置信参考知识图谱作为评估基准，并从实体与关系的覆盖率、准确率等多个维度对初始图谱进行量化评估。再基于当前知识图谱结合大模型生成题目，并由更强的大语言模型从知识覆盖、语义一致性以及实体–关系正确性等多个方面对题目质量进行评估，同时给出针对题目的结构化修改反馈。最后，上述反馈进一步被回溯用于更新初始知识图谱，得到修正后的图谱，并进入新一轮的图谱评估、题目生成与质量评估过程，从而完成**SEKG-QG** 框架的一次完整自进化迭代。实验结果表明，该反馈驱动机制能够显著提升知识图谱质量与题目生成质量，并在多项自动出题任务的测试中证实了该框架在**AQG** 中的稳定性与鲁棒性。因此，**SEKG-QG** 将大语言模型与知识图谱有机结合，为实现高质量且稳定性强的**AQG** 任务提供了一种通用框架。

1 引言

自动出题（Automatic Question Generation, AQG）被广泛应用于教学测评和日常教育应用中，其目标是将教学文档自动转化为可用于评估学习效果的问题集合 [Heilman and Smith, 2010, Kurdi et al., 2020, Bikaun et al., 2024]。通过自动生成高质量的问题，AQG 系统能够在大规模教学场景下减少人工出题成本，并支持持续性的学习评估。与摘要生成或对话生成等任务不同，AQG 对知识正确性和稳定性有更高要求：一旦题目中包含事实错误或关键信息缺失，不仅会降低系统的可信度，还可能误导学习者，从而对教学效果产生负面影响 [Kurdi et al., 2020, Zhou et al., 2023]。

近年来，大语言模型显著提升了自动出题的语言流畅性和多样性 [Raffel et al., 2020, Lewis et al., 2020, Brown et al., 2020]。然而，在实际应用中我们观察到，现有AQG 系统生成的题目质量往往不稳定，不仅在不同学科之间表现差异明显，甚至在同一输入条件下的多次生成过程中，也可能出现题目质量的大幅波动。许多方法采用“先抽取知识、再生成题目”的静态流程：系统从原始文本中抽取实体与关系，构建固定的知识表示，然后基于该知识表示通过大语言模型进行自动出题。但一旦前期知识抽取出现偏差，这些错误便会在后续生成过程中持续放大，而现有系统缺乏有效手段对其进行修正。类似的误差累积现象已在知识图谱构建和知识驱动生成任务中被广泛报道 [Pan et al., 2017, Zhang et al., 2018, Pan et al., 2020]，并进一步影响生成内容的知识一致性和可靠性 [Wang et al., 2022]。

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

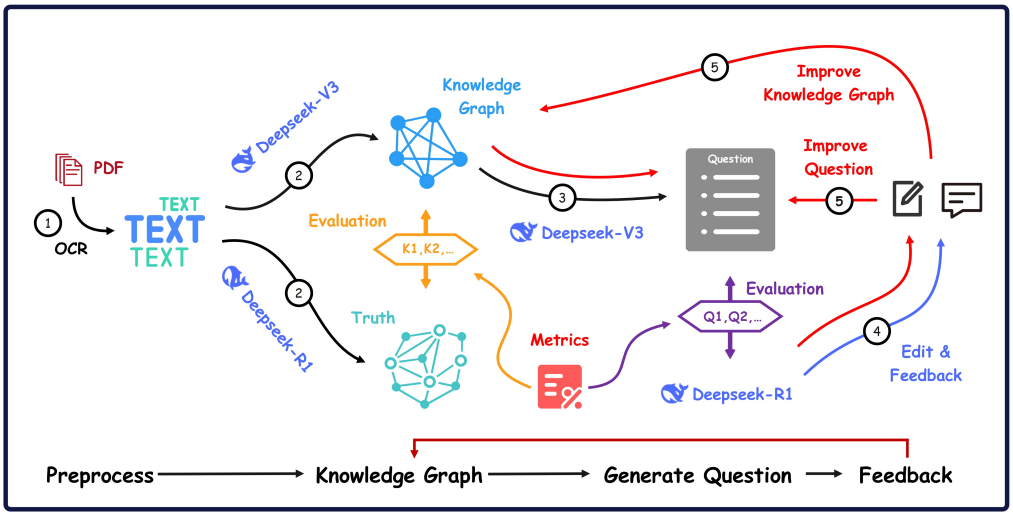

Figure 1: 基于反馈驱动的**SEKG-QG** 框架的技术路线图。系统从PDF 文档中抽取文本段落，结合**DeepSeek-V3** 大模型构建初始知识图谱。进一步使用更强的大模型（**DeepSeek-R1**）构建高置信参考知识图谱结合指标对初始图谱进行评估。随后基于初始知识图谱生成初始问题，并经过DeepSeek-R1 结合指标进行质量评估，同时得到对初始问题的修改和反馈。最后这些修改和反馈反过来修正知识图谱，形成迭代闭环，使知识图谱与生成题目的质量能够协同提升（注意：序号表示框架执行的顺序，迭代闭环结构参见图 1 中红色箭头与紫色评估双向箭头的结合，顺序依次是序号3、紫色评估双向箭头、序号4、序号5，然后再次从序号3执行，开启新一轮的迭代）。

造成上述问题的一个关键原因在于，现有AQG 系统通常缺乏显式且可迭代的质量评估与反馈机制。在缺少有效反馈回路的情况下，系统难以判断当前生成的问题是否覆盖了关键信息，或是否与底层知识表示保持一致，从而导致知识层面的错误在之后的题目生成中不断累积。尽管近期研究探索了基于大语言模型的自反思或自改进生成策略 [Bai et al., 2022, Madaan et al., 2023, Shinn et al., 2023]，但这些方法主要关注对生成文本本身的修正，而未显式更新底层的结构化知识表示，因此难以从根本上缓解知识不一致问题，尤其是在AQG 这类对知识结构高度敏感的任务中。

为此，本文提出一种反馈驱动的知识图谱自进化题目生成框架（**SEKG-QG**），其整体技术路线如图 1 所示。该框架的核心思想是构建一个可评估、可迭代的闭环过程，使知识表示与题目生成质量能够相互促进、协同演化。具体而言，系统以教学常用文档（如PDF）为输入，首先从文本中抽取实体与关系，并结合大语言模型构建初始知识图谱。随后，引入更强的大语言模型构建高置信参考知识图谱，并从实体与关系层面对初始图谱进行量化评估。在此基础上，系统基于当前知识图谱生成题目，并对生成结果进行多维度质量评估，包括知识覆盖性、语义一致性以及实体与关系的正确性，同时输出结构化反馈。这些反馈不仅用于修正生成的问题，还会被显式回溯至知识图谱层，用于更新与纠正原有知识表示，从而形成“知识评估—问题生成—质量反馈—知识修正”的迭代闭环。通过不断重复该过程，**SEKG-QG** 能够逐步提升知识图谱质量，并在此基础上生成更加稳定且高质量的问题。在多种教学文档上的实验结果表明，该框架在提升知识表示与题目质量的同时，有效降低了生成结果的随机性，验证了其在AQG 任务中的稳定性与通用性。

本文的主要贡献如下：

- 提出一种反馈驱动的自进化知识图谱自动出题框架（**SEKG-QG**），通过闭环反馈机制实现知识表示的持续修正与演化；
- 设计面向知识图谱与题目生成的双层质量评估与结构化反馈机制，使题目质量改进能够显式作用于知识结构更新；
- 在多种教学文档上的实验结果表明，该框架在知识表示质量、题目生成质量及生成稳定性方面均显著优于现有基线方法。

代码已开源：https://github.com/undoubtable/KG_allprocess。

2 相关工作

2.1 自动出题与知识建模局限

AQG 是自然语言处理领域的重要研究方向之一，广泛应用于教学测评和日常教育应用中。随着大语言模型（Large Language Model, LLM）的快速发展，AQG 逐渐被建模为序列到序列生成任务，并在大规模数据上实现端到端学习，显著提升了生成问题的语言流畅性与多样性 [Raffel et al., 2020, Lewis et al., 2020, Brown et al., 2020]。

然而，现有AQG 方法在知识建模与一致性保障方面仍存在明显不足。多数方法采用单次生成（single-pass）范式，并依赖静态的知识抽取与表示流程。一旦底层知识表示存在噪声或遗漏，生成模型往往缺乏有效的检测与修正能力，从而在知识密集场景中产生事实不一致或不可靠的问题 [Pan et al., 2020, Wang et al., 2022]。这表明，单纯依赖端到端生成模型，难以满足AQG 任务对知识准确性与生成稳定性的要求。

2.2 基于知识图谱的题目生成

为缓解端到端生成模型在知识建模方面的不足，已有研究将知识图谱等结构化知识表示引入自动出题过程。知识图谱通过实体及其关系对知识进行显式建模，为下游问题生成提供结构化且具有可解释性的约束。相关工作表明，引入知识图谱能够在一定程度上提升生成问题的相关性与事实正确性 [Zhang et al., 2018, Chen et al., 2021]。

然而，多数基于知识图谱的AQG 方法通常假设知识图谱在构建完成后是静态且可靠的。但在实际应用中，知识图谱往往通过自动抽取或弱监督方式获得，不可避免地包含噪声、遗漏或结构性偏差。由于缺乏利用生成结果对知识图谱进行回溯修正的机制，这些知识表示层面的错误可能在出题阶段被直接继承，甚至进一步放大。因此，仅引入结构化知识表示并不足以从根本上解决问题，有效的AQG 系统仍需要一种能够持续精炼知识表示的机制。

2.3 反馈驱动学习与知识精炼

反馈驱动的学习范式在机器学习领域已得到广泛研究，典型方法包括自训练、强化学习以及基于人类反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）等。已有研究表明，引入反馈信号能够有效提升生成模型在对齐性、可靠性与整体可用性方面的表现 [Ouyang et al., 2022, Bai et al., 2022]。

另一方面，人机协同方法也被应用于知识图谱的构建、精炼与验证过程中，通过引入人工反馈提升知识获取与表示的质量 [Bikaun et al., 2024]。尽管这些方法强调反馈在知识质量控制中的重要作用，但往往高度依赖显式人工监督，不仅成本较高，也难以在大规模或多学科场景中推广。此外，现有工作通常将反馈机制与下游生成任务分离，尚缺乏一种系统化方式，能够将生成过程中产生的反馈直接用于驱动显式知识表示的迭代精炼。

2.4 知识图谱与大型语言模型

随着LLM 的快速发展，越来越多的研究开始探索将知识图谱（Knowledge Graph，KG）与LLM 相结合，以增强模型的推理能力、事实一致性与生成可控性 [Yasunaga et al., 2021, Yao et al., 2023, Sun et al., 2023]。在这些工作中，KG 通常被用作外部记忆或结构化引导，从而为生成过程提供更加有依据且具备一定可解释性的支持 [Liu et al., 2024, Wang et al., 2024]。相关研究表明，将符号化知识与神经语言模型相结合，在知识密集型任务中具有显著潜力。

然而，多数现有方法仍将KG 视为固定输入，并未显式建模知识表示随生成结果而动态演化的过程。与此同时，LLM 在生成过程中所体现的强语义理解与评估能力，也很少被进一步用作结构化反馈信号以更新底层KG。与以往工作不同，本文聚焦于构建一种反馈驱动的闭环框架：LLM 不仅作为问题生成器与质量评估器，同时作为结构化反馈的来源，驱动KG 的持续演化，从而协同提升知识表示质量与自动出题性能。

3 问题定义

我们研究一种基于反馈驱动的知识图谱自进化题目生成框架（Fig. 1），其核心目标是在迭代过程中，利用LLM 产生的结构化反馈持续提升KG 的质量，并生成高质量的问题集合。

给定一组常用教学文档（如PDF）：

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

对于每个文档d_i，首先通过Optical Character Recognition（OCR）提取文字，随后逐页将其划分为文本段落集合：

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

基于文本使用选定的基准LLM 构建初始知识图谱$G^{(1)} = (E^{(1)}, R^{(1)})$，其中$E^{(1)}$ 表示抽取的实体集合，$R^{(1)}$ 表示实体之间的关系集合。第t 次迭代的知识图谱记为$G^{(t)} = (E^{(t)}, R^{(t)})$。

为量化评估知识表示质量，我们假设可以借助更强的LLM 构建高置信参考知识图谱（基准）$G^* = (E^*, R^*)$，并通过一组图谱质量指标$\mathcal{M}_K^{(t)}$（如实体覆盖率、关系覆盖率以及实体、关系正确性等）对$G^{(t)}$ 进行评估。

随后，基于$G^{(t)}$ 及原始文本段落，基准LLM 生成题目集合$Q^{(t)}$。这些题目随后由更强的LLM 进行质量评估（题目质量指标$\mathcal{M}_Q^{(t)}$）并给出相应的题目修改反馈信号$\mathcal{F}^{(t)}$。

本文关注的核心问题是，如何有效利用反馈信号$\mathcal{F}^{(t)}$ 对$G^{(t)}$ 进行迭代更新，即实现

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

从而使知识表示质量与自动出题性能在迭代过程中持续协同提升。

4 SEKG-QG 框架

我们提出的**SEKG-QG** 框架用于用于自动出题任务中知识表示与题目生成性能的联合优化。该框架采用反馈驱动的迭代闭环流程，依次包含知识图谱构建、基于知识图谱的问题生成、基于大语言模型的质量评估，以及反馈驱动的知识图谱自进化。通过多轮迭代，底层知识表示得以持续修正与演化，从而逐步提升自动出题的稳定性与整体质量。

与将知识图谱视为静态输入的现有方法不同，**SEKG-QG** 显式利用下游问题评估阶段产生的结构化反馈，对知识图谱进行持续修正与改进，从而形成以反馈为核心的闭环演化系统。

4.1 知识图谱构建

给定从文档中抽取的文本段落集合，我们采用大语言模型**DeepSeek-V3** 进行实体识别与关系抽取，构建初始知识图谱：

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

其中$E^{(1)}$ 表示抽取的实体集合，$R^{(1)}$ 表示抽取的关系集合。

在该阶段，构建过程优先保证实体与关系的覆盖率而非精确率，因此允许有意引入一定噪声。这些噪声预计将在后续迭代中，通过反馈驱动的知识精炼过程被逐步纠正。

为实现可靠的质量评估，我们进一步采用更强的大语言模型，并结合更保守的抽取提示，构建高置信参考的知识图谱（基准）：

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

需要强调的是，G^* 并非人工标注的真值，而是作为知识图谱质量评估的基准。

4.2 知识图谱评估指标

在第t 次迭代中，当前的知识图谱记为：

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

为对知识图谱质量进行量化评估，我们在宽松匹配（relaxed matching）设定下，定义一组知识层面的评估指标$\mathcal{M}_K$ [Paulheim, 2017, Hogan et al., 2021]。该设定允许在一定程度上忽略词汇表述差异，更好地适配基于LLM 的知识抽取场景，从而合理刻画实体与关系在语义层面的对应关系。

实体相关指标　实体精确率与召回率分别刻画实体抽取的正确性与覆盖完整性：精确率通过惩罚冗余或错误实体反映抽取质量，而召回率通过惩罚遗漏实体反映知识覆盖程度。

- 实体精确率（**Entity Precision**）

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

- 实体召回率（**Entity Recall**）

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

  在宽松匹配设定下，实体召回率亦可视为实体覆盖率。

- 实体**F1**（**Entity F1**）

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

关系相关指标

- 关系精确率（**Relation Precision**）

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

- 关系召回率（**Relation Recall**）

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

  在宽松匹配设定下，关系召回率亦可视为关系覆盖率。

- 关系**F1**（**Relation F1**）

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

汇总　实体层面与关系层面的指标共同刻画知识图谱的实体与关系正确性。在实验中，我们分别报告各项指标以进行细粒度分析；同时也可将其聚合为标量形式的知识质量得分$K^{(t)}$，用于刻画不同迭代轮次中的整体改进趋势 [Paulheim, 2017]。

4.3 基于知识图谱的题目生成

在每次迭代中，系统基于当前知识图谱$G^{(t)}$与原始文本段落生成一组问题：

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

该阶段的目标是在保持问题与知识表示结构一致性的前提下，最大化对知识图谱中实体与关系的覆盖。因此，该阶段采用规模适中的大语言模型**DeepSeek-V3**，以在生成质量与计算效率之间取得平衡，而非依赖最强的可用模型。

4.4 问题质量评估指标

生成的问题集合$Q^{(t)}$将由更强的大语言模型**DeepSeek-R1**进行自动评估，该模型在框架中充当评估器（evaluator） [Liu et al., 2023, Zheng et al., 2023b]。给定当前知识图谱$G^{(t)} = (E^{(t)}, R^{(t)})$，评估器从多个互补维度对每个问题$q \in Q^{(t)}$进行质量评估。

题目质量相关指标　我们定义4个归一化到$[0,1]$的题目质量相关指标如下（$\mathcal{E}_q$与$\mathcal{R}_q$分别表示出现在题干与选项中的实体与关系集合。）：

- 知识覆盖度（**Knowledge Coverage,** A）衡量单道题目所涉及的实体与关系，在当前知识图谱中的覆盖比例，反映问题对底层知识表示的利用程度：

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

- 语义连贯性（**Semantic Coherence,** B）衡量题干与选项在语义和逻辑上的一致性，由评估器根据问题表述是否清晰、选项是否合理给出评分 [Liu et al., 2023]：

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

- 实体对齐准确率（**Entity Alignment Accuracy,** C）衡量问题中出现的实体能否被正确对齐到当前知识图谱中，从而反映问题在实体层面与知识表示的一致性。其中，$\mathcal{E}_q^{\text{align}} \subseteq \mathcal{E}_q$表示在问题$q$中出现的实体被评估器判定为能够正确对齐到$E^{(t)}$中某一实体的子集：

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

  其中$\epsilon > 0$为一个足够小的常数，用于数值稳定性。

- 关系正确性（**Relation Correctness,** D）衡量题目所蕴含的实体间关系，是否与知识图谱中定义的关系保持一致，从而反映问题在关系层面的事实正确性。其中，$\mathcal{R}_q^{\text{corr}} \subseteq \mathcal{R}_q$表示由问题正确答案所蕴含，且被评估器判定为与$R^{(t)}$中关系语义一致的关系子集：

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

上述指标加权聚合得到总体的题目质量得分：

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

其中w_A, w_B, w_C, w_D为非负权重，本文默认采用均匀权重$w_A = w_B = w_C = w_D = 0.25$。评估器输出题目质量的量化指标$\mathcal{M}_Q^{(t)}$的同时，也生成了结构化文本反馈信号：

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

用于指出实体或关系缺失、表述歧义等问题，并指导下一轮迭代中的知识图谱更新。

整套题目评估　除对单道题目进行评估外，我们进一步从全局角度对整套题目$Q^{(t)}$进行评估，以衡量其在实体与关系层面的覆盖分布是否合理。在自动出题场景中，若题目集中在少数实体或关系上，容易导致知识覆盖偏置，从而削弱题集在教学或测评中的代表性与有效性 [Kurdi et al., 2020]。

并据此计算其经验分布$P_{\mathcal{E}}$与$P_{\mathcal{R}}$。我们进一步定义目标分布$P_{\mathcal{E}}^*$与$P_{\mathcal{R}}^*$，用于刻画理想情况下题目对实体与关系的覆盖分布（例如均匀分布，或由知识图谱结构诱导的先验分布）。

整套题目的结构质量定义为经验分布与目标分布之间的匹配程度：

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

其中$\text{Dist}(\cdot)$表示对称的分布散度度量，本文采用Jensen–Shannon 散度以保证度量的有界性与稳定性 [Lin, 1991]。$S(Q^{(t)})$越高，表示题集在实体与关系层面的覆盖分布越接近目标分布，从而在结构上更加均衡、覆盖更加充分。

4.5 反馈驱动的知识图谱自进化

在每一轮迭代中，评估阶段产生的反馈信号$\mathcal{F}^{(t)}$会被解析为结构化的知识精炼操作，包括实体纠错、关系修正以及缺失知识补全等。这些操作作用于当前知识图谱，得到更新后的图谱：

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

其中$\Phi(\cdot)$表示反馈驱动的知识更新算子。更新后的知识图谱$G^{(t+1)}$将进入下一轮迭代，继续用于基于**DeepSeek-V3** 的问题生成，并由**DeepSeek-R1** 进行自动评估。通过多轮迭代，知识图谱逐步向更高覆盖率与更高正确性演化，从而带来自动出题性能的持续提升。

5 实验

5.1 实验设置

我们在一组教学测评类PDF 文档（由一套法学专业教材拆分的40个PDF文档）上评估所提出的**SEKG-QG** 框架。每个PDF 文档首先被解析为文本段落，作为知识图谱构建与问题生成的输入。

对于每个文档，我们比较**SEKG-QG** 处理前后的两种状态：由初始抽取得到的知识图谱$G^{(1)}$及其生成的问题集合$Q^{(1)}$，以及经过一轮反馈驱动自进化后的知识图谱$G^{(2)}$ 与对应的问题集合$Q^{(2)}$。除是否进行反馈驱动精炼外，其余实验设置保持一致。

5.2 评估指标

我们从知识表示与问题生成两个层面对方法效果进行评估。

知识图谱指标 知识图谱质量通过实体与关系层面的精确率、召回率与F1 分数进行评估，并以由DeepSeek-R1 构建的高置信参考知识图谱作为对照。相关指标定义见第 4.2 节，聚合得分记为$K^{(t)}$。

问题生成指标 问题生成质量由DeepSeek-R1 作为自动评估器进行评估，相关指标定义见第 4.4 节。题目质量的最终得分$Q^{(t)}$ 通过公式18 计算得到，并基于公式20 进一步计算整套题的得分$S(Q^{(t)})$，以衡量题集在实体与关系层面的覆盖均衡性。

5.3 知识图谱质量对比

我们首先分析SEKG-QG 框架使用前后对知识图谱质量的影响，对40个PDF文件经过SEKG-QG框架处理前后的图谱质量进行统计，这里为了凸显对比只选取了迭代1次的结果（见表 1）。可以观察到自进化后的知识图谱在实体与关系两个层面均取得一致提升（自进化后的图谱质量$\mathcal{M}_K^{(2)}$ 均一致优于初始图谱质量$\mathcal{M}_K^{(1)}$）。

其中，实体召回率（覆盖率）从0.792 提升至0.858，表明精炼过程有效补全了初始图谱中的遗漏实体。关系层面的提升尤为显著，关系F1 从0.171 提升至0.320，说明精炼过程能够显著改善关系结构的准确性与完整性。

Table 1: SEKG-QG 框架应用前后的知识图谱质量指标$\mathcal{M}_K^{(t)}$ 对比

方法	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 生成题目质量对比

Table 2: SEKG-QG 框架应用前后的问题质量指标$\mathcal{M}_Q^{(t)}$ 对比

方法	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**

我们依旧是选取上述的40个PDF文件，进一步评估知识图谱自进化对生成问题质量的影响，题目生成的设置是每个PDF文件生成最多50道题目，但是经过实际统计共40套题

目，1672道题目，基于这些数据平均后整理得到表 2的结果。可以看出，尽管在知识覆盖度$A(q)$ 上提升不大，但基于自进化的图谱生成问题在语义连贯性、实体对齐准确率与关系正确性方面均取得显著提升。在综合指标上，应用SEKG-QG框架后题目的平均质量得分$Q(q)$ 由75.33 提升至77.13，整套题目的平均得分$S(Q^{(t)})$ 亦从76.42 提升至79.29，表明生成题集在结构可靠性与覆盖均衡性方面得到明显改善。

5.5 案例分析

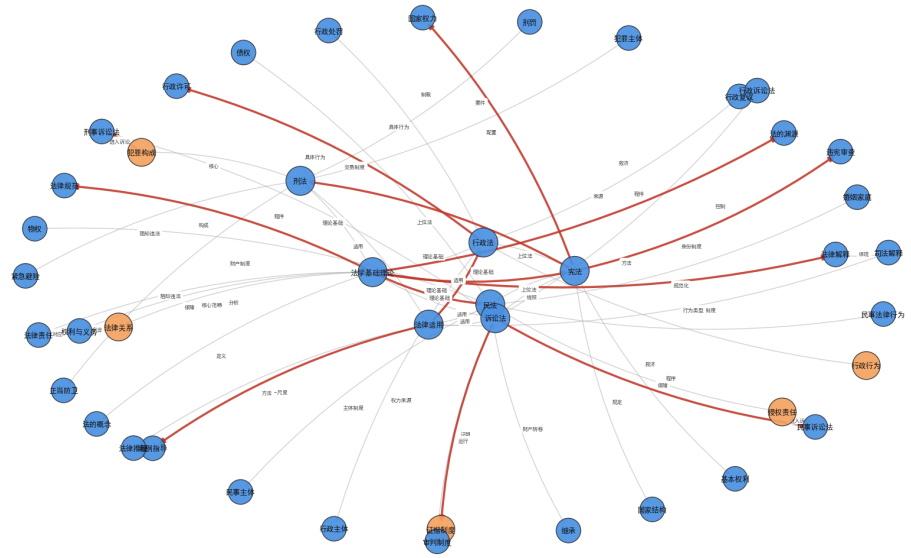

Figure 2: 随机选取的某一知识图谱进化前后的对比，蓝色节点和灰色边表示初始图谱的实体和关系，橙色节点表示知识图谱自进化后新增的节点，红色的边表示纠正后的关系，可以看到SEKG-QG框架确实能提升知识图谱的质量，并且也比较符合表 1的结果，即实体修改较少，关系纠正比较多。

图 2 给出了随机选取一个文档经过SEKG-QG框架处理前后的知识图谱与生成题目的前后对比。可以观察到，初始知识图谱中原有的关系缺失与错误得到修正，相应生成的问题在实体指向与关系表述上更加准确，并与底层知识结构保持更好的一致性。

6 讨论

所提出的SEKG-QG 框架表明：反馈驱动的知识图谱自进化题目生成是一种能够同时提升结构化知识表示与自动出题质量的有效范式。与将知识图谱视为静态中间产物的传统流水线方法不同，我们将知识表示建模为一种可被评估、可被修正、并能够依据生成结果持续演化的动态对象

为何反馈驱动演化有效 实验结果揭示了一个关键现象：出题过程中暴露的错误往往直接反映了底层知识图谱中的结构性缺陷，例如实体遗漏、关系错误或抽象粒度不当。通过引入大型语言模型作为评估器，SEKG-QG 能够将生成阶段显现的问题转化为可操作的结构化反馈，并用于驱动知识图谱的迭代进化 [Paulheim, 2017, Hogan et al., 2021]。。由此，知识图谱与题目生成过程之间形成闭环，系统得以修正自身的结构性知识，而不仅仅停留在生成策略或解码层面的局部调整。

覆盖率与精确性的权衡 实验还表明，知识覆盖率与结构精确性之间存在内在权衡关系。初始知识图谱通常以覆盖为导向，从而不可避免地引入噪声，尤其体现在实体间的关系层面。反馈驱动的图谱自进化过程能够逐步削减这些噪声，在不显著牺牲覆盖范围的前提

下，提升关系精确性与整体一致性。这一结果表明，在具备可靠反馈机制的前提下，早期阶段允许一定噪声以换取更高覆盖率，可能是一种合理且有效的策略。

大型语言模型作为评估器的作用　不同于传统的自动指标或基于规则的验证方法，大型语言模型能够提供更灵活且语义更丰富的评估能力，从而捕捉实体错配、逻辑不一致等细粒度问题 [Liu et al., 2023]。在**SEKG-QG** 中，LLM 不仅充当生成器，同时也作为评估器与反馈提供者，在较低边际成本下近似人工审核过程，并输出可直接用于知识更新的结构化反馈。这一观察进一步表明，将LLM 作为反馈来源用于知识中心系统具有重要潜力。

局限性　尽管**SEKG-QG** 框架取得了稳定改进，仍存在若干局限。首先，依赖较强的LLM 进行评估与参考图谱构建会带来额外的计算成本，且评估器本身可能引入模型偏差 [Zheng et al., 2023a]。其次，参考知识图谱G^* 并不保证是真实的绝对真值，这可能影响部分定量结果的可解释性。此外，当前反馈解析与精炼算子$\Phi(\cdot)$ 仍相对粗粒度，未来可结合置信度建模与不确定性估计，进一步提升知识更新的精细性。

未来方向　未来工作可从以下多个方向扩展该框架：

- 设计自适应停止准则，例如基于知识质量或题目质量的增长收敛情况，自动确定演化迭代轮次；
- 通过评估器蒸馏或轻量化模型，降低基于LLM 的评估与反馈生成成本；
- 更紧密地联合优化知识图谱结构与出题目标，例如将知识覆盖与题目质量指标显式纳入统一优化过程；
- 将反馈驱动的闭环演化为范式，推广至其他知识密集型生成任务，如解释生成、课程内容组织与教育内容推荐等。

总的来讲，该框架可推广至其他依赖知识表达的生成任务，如解释生成与教育内容推荐等。总体而言，SEKG-QG 框架为提升LLM 与结构化知识系统的可靠性与可控性提供了一种具有普适性的机制。

7　结论

本文提出**SEKG-QG**，一种基于反馈驱动的知识图谱自进化生成题目的框架。该框架通过构建“知识图谱构建、知识引导出题、LLM 自动评估、反馈驱动图谱自进化、再次出题”的迭代闭环，实现了知识图谱质量与出题性能的协同、持续提升。

实验结果表明，SEKG-QG 框架的应用能够显著改善实体覆盖、关系正确性与整体结构质量，并进一步带来更可靠、更语义一致的自动出题效果，包括语义连贯性、实体对齐准确率与关系一致性的提升。不同于依赖静态知识表示或单轮生成的现有方法，**SEKG-QG** 提供了一种反馈闭环的机制：利用生成阶段暴露的知识缺陷进行回溯修正，从而持续优化显式知识表示。

更广泛地看，我们的研究表明，将结构化知识视为可演化、可修正的动态组件，而非固定的中间产物，能够显著增强知识密集型生成系统的鲁棒性与可控性。我们相信，所提出的反馈驱动范式为在更广泛应用场景中深度整合大型语言模型与结构化知识提供了坚实而可扩展的基础。

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.