

DOI: 10.3969/j.issn.1671-0673. 2026. 04. 010

面向动态知识检索的大语言模型与知识图谱协同推理方法

夏 毅, 贺浩峰, 陈 静, 周 刚, 郑嘉丽, 张凯翔

(信息工程大学, 河南 郑州 450001)

摘要: 针对大语言模型存在知识静态性及事实幻觉, 而知识图谱又面临覆盖不全与更新滞后的问题, 提出一种面向动态知识检索的大语言模型与知识图谱协同推理框架。该框架通过闭环机制实现知识动态检索、协同推理与增量更新。首先, 采用子图增强的检索生成策略, 基于问题分解与多跳推理从知识图谱中提取相关子图, 提升复杂查询下的知识召回精度; 其次, 通过多维度混合加权方法, 从非结构化文本中检索相关段落以补充知识缺失; 最后, 借助大语言模型对多源知识进行融合推理, 生成答案并抽取出新增三元组, 实现知识图谱的动态增量更新。实验表明, 该方法能有效处理复杂多跳问题, 显著降低事实幻觉, 并通过闭环更新保障知识库的时效性与持续演进能力。

关键词: 知识推理; 大语言模型; 知识图谱; 检索增强生成

中图分类号: TP391.3

文献标识码: A

文章编号: 1671-0673(2026)04-0450-07

Collaborative Reasoning Method Between LLMs and Knowledge Graphs via Dynamic Knowledge Retrieval

XIA Yi, HE Haofeng, CHEN Jing, ZHOU Gang, ZHENG Jiali, ZHANG Kaixiang

(Information Engineering University, Zhengzhou 450001, China)

Abstract: To address the issues of static knowledge and factual hallucinations in large language models (LLMs), as well as the incomplete coverage and delayed updates in knowledge graphs, a collaborative reasoning framework integrating LLMs and knowledge graphs is proposed via dynamic knowledge retrieval. A closed-loop mechanism is designed to implement dynamic knowledge retrieval, collaborative reasoning, and incremental updating. Firstly, a subgraph-enhanced retrieval strategy is adopted, which extracts relevant subgraphs from the knowledge graph through question decomposition and multi-hop reasoning to improve the recall under complex queries. Secondly, semantically relevant passages are retrieved from unstructured texts through a multi-dimensional hybrid weighting method to supplement missing knowledge. Finally, LLMs are utilized to integrate multi-source knowledge for reasoning, generating answers while extracting new triples for dynamically updating the knowledge graph. Experiments demonstrate that the proposed method can effectively handle complex multi-hop questions, reduce factual hallucinations, and ensure timeliness and continuous evolution of the knowledge base through closed-loop updating.

Key words: knowledge reasoning; LLMs; knowledge graph; retrieval-augmented generation

收稿日期: 2026-01-26; 修回日期: 2026-05-08

基金项目: 国家自然科学基金(42371438); 河南省科技攻关项目(222102210081, 222300420590)

作者简介: 夏 毅(1997—), 男, 博士生, 主要研究方向为知识的表示与推理。E-mail: summer_one520@163.com

通信作者: 张凯翔(1990—), 男, 讲师, 硕士, 主要研究方向为大语言模型。E-mail: kaixiang215@126.com

近年来,以GPT(Generative Pre-Training)系列为代表的大语言模型(Large Language Models, LLMs)在自然语言理解、生成与推理任务中展现出卓越性能,推动了人工智能技术在各领域的广泛应用。然而,大语言模型依赖其预训练语料中的静态知识,难以捕捉现实世界的动态变化,且在生成过程中易出现“事实幻觉”问题,这严重限制了其在金融、医疗等高可靠性场景中的应用^[1]。

为提升大语言模型的知识可靠性与时效性,研究者开始探索外部知识增强路径。知识图谱(Knowledge Graphs, KG)作为一种用图结构描述实体及其关系的语义网络,存储了大量结构化的事实知识。与大语言模型的参数化隐性知识不同,知识图谱的知识是显性的、符号化的,具有高精度和良好的可解释性^[2]。将知识图谱融入大语言模型的推理过程,可为模型提供可靠的事实依据,从而缓解幻觉问题^[3]。

然而,传统的知识图谱存在其固有的局限性。一方面,尽管现有知识图谱(如Wikidata、DBpedia)规模庞大,但相较于浩如烟海的人类知识,其覆盖范围仍然有限,存在信息大量缺失的问题。另一方面,知识图谱的构建和更新成本高昂,难以跟上知识爆炸式增长的速度,导致其内容同样存在滞后性。若定期由领域专家检查和验证知识图谱,以确保准确性和相关性,则其构建与维护的成本高昂且难以适应知识的快速演进。

与此同时,互联网中海量的非结构化文本构成了最鲜活、最广泛的知识来源^[4],通过信息检索系统可以实现相对完备知识的获取,但其信息呈碎片化分布,缺乏内在关联与语义组织。因此,如何有机融合大语言模型的泛化推理能力、知识图谱的结构化表示能力以及非结构化文本的动态覆盖能力,已成为构建下一代可靠、动态、可解释知识系统的关键挑战。现有研究多侧重于单向的知识注入或静态检索增强,尚未在机制层面实现大语言模型、信息检索系统和知识图谱三者推理过程中的实时协同与闭环演化。

从技术范式来看,三者形成了功能互补的“黄金三角”。大语言模型以生成模型为基础,具备强大的语义理解与泛化能力,但存在事实可靠性不足的问题^[5];信息检索系统基于检索模型,能够提供可靠的事实依据,但难以处理需要复杂推理的问答^[6];知识图谱通过结构化表示显式建模实体关系,可提升复杂问答的推理能力,但其知识覆盖面

与时效性受限^[7]。这种互补性为构建更强大的知识系统提供了理论基础,但如何实现三者的有机协同仍是一个开放问题。

针对上述挑战,图检索增强生成(Graph Retrieval-Augmented Generation, GraphRAG)范式应运而生,成为融合知识图谱与大语言模型的主流技术路径,近年来在检索机制与知识更新方面取得显著进展。在检索机制优化方面,ToG-2(Think-on-Graph 2.0)^[8]提出实体链接驱动的文档—图谱交替检索框架,通过紧耦合策略提升多跳推理性能,但其知识库仍为静态,缺乏动态演进能力。LightRAG^[9]采用图索引与向量检索的混合架构,支持本地/全局/混合多策略检索,在工程灵活性上具有优势,但未涉及检索结果的知识反馈与图谱更新。在闭环知识演化方向,RAG-KG-IL^[10]引入多智能体增量学习机制,支持动态知识适配而无需全量重训练,有效降低幻觉率,但其更新焦点集中于模型参数微调,而非知识图谱本身的结构演进。WTS(Way-to-Specialist)^[11]提出“LLM $\cup$ KG”双向增强范式,通过领域知识迭代提升专业化能力,验证了闭环演进的可行性,但其目标侧重于领域适配,与本文关注的通用知识图谱时效性维护存在着定位差异。

本文提出一种面向动态知识检索的协同推理框架(Synergistic Knowledge Graph RAG, SKG-RAG),旨在构建一个动态、闭环的知识增强系统。该框架的核心在于将动态知识检索过程深度融入协同推理机制。在响应用户查询时,系统通过多跳问题分解,从知识图谱中检索相关子图,同时从非结构化文档中匹配补充段落;利用大语言模型对多源异质知识进行融合推理,生成可靠答案;更重要的是,系统能够从高质量的推理结果中自动抽取出新的结构化知识,并将其动态反馈至知识图谱,实现系统的自我演化和持续成长。该研究为构建具备持续学习能力的可信人工智能系统提供了新思路,在需要动态知识更新的智能问答、决策支持等场景具有重要应用价值。

1 协同知识推理方法

1.1 整体流程

SKG-RAG是一个基于子图检索增强生成的混合知识问答系统,其架构深度融合了结构化知识(知识图谱)与非结构化知识(文本语料),并通过大语言模型进行统一调度与推理。该框架将大语

言模型的语义理解与推理能力,与结构化、非结构化知识的动态检索过程紧密结合,在处理多跳推理、时序敏感及知识缺失等复杂场景时表现出显著优势。在面向动态知识演化的推理系统中,知识图谱作为结构化的核心知识载体,其完整性、准确性与时效性直接决定系统的推理质量。然而,知识图谱本质上是某一时刻知识的静态快照,难以覆盖开放域中持续涌现的新事实,亦无法自动识别已有三元组是否因现实世界演变而失效(例如机构隶属变更、人物职务更替、政策法规更新等)。若仅依赖知识图谱单独响应用户查询,系统在面临信息缺失或知识过时的情况下,容易返回空结果、陈旧答案或低相关性子图,导致推理失败或结论可信度下降^[12]。

同时,大语言模型本身缺乏明确的时间感知与事实核查机制,其训练数据具有时间截断性,且输出易受提示方式与上下文偏差干扰,导致对“知识是否过时”的判断主观性强、可靠性不足。若基于此类不可靠判断直接更新知识图谱,可能引入

事实幻觉或将错误信息固化,反而损害知识库的一致性、权威性与可维护性。传统问答系统若单独依赖大语言模型,易受其知识静态性、事实幻觉与推理不透明等局限影响^[13];而仅依靠知识图谱或文本检索,则难以实现跨域、多层次语义关联与逻辑推理。

本文提出的SKG-RAG框架是一种以用户查询为驱动、以推理过程异常为触发信号的动态更新机制,如图1所示。该机制的核心思想是:不预设知识图谱是否完整或过时,而是在实际问答过程中,通过知识感知实时评估当前知识图谱能否提供有效、完整、高置信的推理支持。当系统检测到知识不足、证据矛盾或置信度过低时,将自动触发动态知识检索与多源融合流程。其整体工作过程可描述为:针对自然语言问题,系统自适应地从知识图谱中检索语义与结构相关的子图,并从非结构化文档库中匹配相关文本段落,进而将多源异构知识融合为高质量上下文,供大语言模型进行协同推理并生成最终答案。

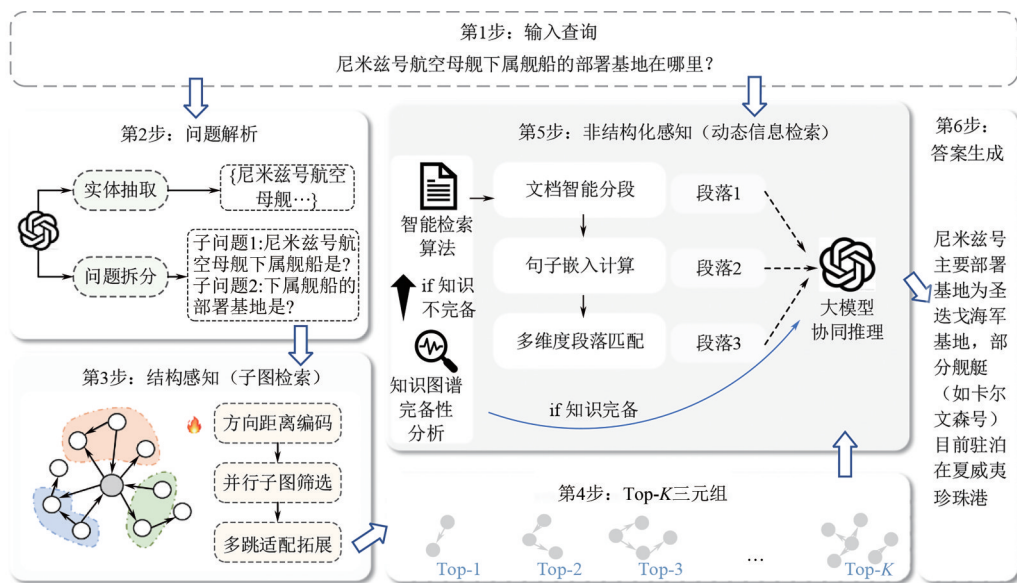


图1 SKG-RAG模型框架图

所提框架各流程具有以下4个核心特征。1)多源知识融合。协同利用知识图谱的结构化语义网络与非结构化文本的动态知识覆盖,构建互补增强的知识支撑体系。2)动态混合检索。通过问题语义分解、实体链接与多跳推理,从知识图谱中动态抽取相关子图,并结合多维度匹配算法从文本库中检索补充段落,实现精准、高效的知识获取。3)大语言模型协同推理。以大语言模型作为中央调度与推理引擎,统筹完成问题解析、检索策略生成、多源知识融合、置信度评估与最终答案生

成。4)闭环知识演化。系统可从高质量推理结果中自动抽取高置信度的新知识三元组,经校验后增量更新至知识图谱,实现知识库的持续自主演进与时效性维护。

1.2 问题解析

针对复杂问题隐含的多层语义与实体依赖^[14],SKG-RAG首先利用大语言模型对问题进行分解。通过识别核心实体与逻辑关系,将原始问题拆解为多个逻辑递进的子问题,并显式记录其间的依赖关系,以指导后续结构化检索与多源信息融合,

从而保障推理路径的连贯与完整。

1.3 结构感知:子图检索

针对传统RAG检索粒度粗、无法感知知识结构的问题,设计一种结构感知的子图检索算法,旨在从海量三元组中高效筛选出与问题最相关的证据子图,为推理提供精准、可解释的结构化知识支撑。该算法的核心优势在于融合语义特征与结构特征,将检索问题转化为可并行的三元组分类任务,具体流程如下。

首先,构建与问题相关的Topic实体集。对拆解后的每个问题 q ,通过查询字符串进行分词、小写化、标点移除后,与知识图谱实体集 E 进行最大前向匹配,生成初始候选实体集合 C_q 。随后利用预训练文本编码器分别编码子问题与候选实体,并通过余弦相似度从候选实体集合 C_q 筛选出与问题语义相关的主题实体,表示为

$$T_q = \{e \in C_q \mid \text{sim}(e, q) > \tau\} \quad (1)$$

式中: e 表示候选实体; $\text{sim}(\cdot)$ 表示余弦相似度函数; τ 表示相似度阈值,实验中设为0.6; T_q 表示Topic实体集,选取相似度超过阈值 τ 的实体作为主题实体,作为子图检索的中心种子。

为捕捉实体间的多跳关联与方向性关系,采用多轮结构传播策略,对Topic实体周围的三元组进行定向传播。首先,对Topic实体集中的实体构造初始指示状态,即

$$s_e^{(0)} = \begin{cases} 1, e \in T_q; \\ 0, e \notin T_q. \end{cases} \quad (2)$$

式中, $s_e^{(0)}$ 表示实体 e 在第0轮传播时的初始结构状态。随后在知识图谱 G 中执行前向传播与反向传播,通过方向性距离编码量化实体间的结构关联强度。其前向更新公式为

$$s_e^{(\ell+1)} = \text{MEAN} \{ s_{e'}^{(\ell)} : (e', r, e) \in G \} \quad (3)$$

反向更新公式为

$$s_{e_{\text{rev}}}^{(\ell+1)} = \text{MEAN} \{ s_{e'}^{(\ell)} : (e, r, e') \in G \} \quad (4)$$

式中: $\ell = 0, 1, \dots, L-1$,初始条件为 $s_{e_{\text{rev}}}^{(0)} = s_e^{(0)}$; r 表示关系类型; $\text{MEAN}(\cdot)$ 表示均值聚合函数。前向传播用于聚合以 e 为尾实体的入边邻居实体 e' 的状态,反向传播用于聚合以 e 为头实体的出边邻居实体 e' 的状态。两者结合可近似刻画实体到Topic实体集合 T_q 的方向性结构距离,从而为后续三元组打分提供结构特征。

将多轮前向传播和反向传播得到的状态进行拼接,形成实体 e 的最终结构编码向量,表示为

$$s_e = [s_e^{(0)} \parallel s_{e_{\text{rev}}}^{(1)} \parallel \dots \parallel s_e^{(L)} \parallel s_{e_{\text{rev}}}^{(1)} \parallel \dots \parallel s_{e_{\text{rev}}}^{(L)}] \in \mathbb{R}^{2L+1} \quad (5)$$

式中: s_e 表示实体 e 的结构编码向量; $\parallel$ 表示拼接操作。

在此基础上,利用预训练文本编码器分别计算实体文本、关系文本及问题文本的语义嵌入,记为 z_e, z_r 和 z_q 。在距 T_q 不超过 R 跳的候选子图内,对每条三元组 $\tau = (h, r, t)$,构造联合特征为

$$x_\tau = [z_q \parallel z_h \parallel z_r \parallel z_t \parallel s_h \parallel s_t] \quad (6)$$

式中: z_h 表示头实体 h 的语言嵌入; z_t 表示尾实体 t 的语言嵌入; s_h 表示头实体 h 的结构编码向量; s_t 表示尾实体 t 的结构编码向量。

随后,将其输入含单隐藏层的多层感知机(Multi-Layer Perceptron, MLP),经Sigmoid函数归一化后得到三元组 τ 在给定问题 q 条件下的保留概率:

$$p_\theta(\tau|q) = \sigma(\text{MLP}_\theta(x_\tau)) \quad (7)$$

式中: $\sigma(\cdot)$ 表示Sigmoid函数; θ 表示可训练的参数集合。

按降序选取前 K 条三元组,构成子图:

$$G_q = \text{Top-}K(p_\theta(\tau|q)), \tau \in C_q \quad (8)$$

1.4 非结构化感知:动态信息检索

为补充知识图谱的覆盖不足并满足动态信息需求,设计动态外部知识检索模块,采用多维度混合加权检索算法,从外部文档中获取相关知识。该算法不局限于单一语义相似度,而是融合语义、实体和关键词3个维度进行协同匹配,既避免单一策略的片面性,又能复用已检索的结构化知识,从而提升非结构化知识的检索精度与鲁棒性,为后续推理提供全面支撑。

在语义维度上,采用预训练句子嵌入模型 $\mathcal{M}$ 将用户问题 q 与文档段落 s_i 映射到统一高维向量空间,分别得到语义嵌入 q_{emb} 和 $s_{i,\text{emb}}$ 。通过余弦相似度量二者语义关联强度,其计算公式为

$$\text{sim_sem}(q, s_i) = \frac{q_{\text{emb}} \cdot s_{i,\text{emb}}}{\|q_{\text{emb}}\| \cdot \|s_{i,\text{emb}}\|} \quad (9)$$

实体维度的设计核心是利用结构化领域关联信息进行筛选,本文直接复用子图检索阶段得到的Topic主题实体集 T_q ,量化段落 s_i 与问题 q 的领域关联性,其计算公式为

$$\text{score_en}(s_i) = \min \left(\sum_{e \in T_q} w_e \cdot I(e \in s_i), 3 \right) \quad (10)$$

式中: w_e 表示实体匹配的权重; $I(\cdot)$ 表示指示函数。

关键词维度聚焦于捕捉问题中的关键信息,弥补语义维度对精准术语、失效词等细节信息的

敏感不足的问题,通过TF-IDF算法从用户问题 q 中提取核心关键词集合 $K_q = \{k_1, k_2, \dots, k_m\}$,统计段落中的关键词命中比例,计算公式如下:

$$\text{score_key}(s_i) = \frac{1}{|K_q|} w_k \cdot I(k \in s_i) \quad (11)$$

式中, w_k 为关键词匹配的权重。

段落 s_i 的最终相关性得分通过加权求和得到:

$$\text{score_fn}(s_i) = \text{sim_sem}(q, s_i) + \text{score_en}(s_i) + \text{score_key}(s_i) \quad (12)$$

通过对上述3个维度进行加权融合,系统能够在抑制噪声段落的同时,兼顾语义相关性、领域一致性与信息精确性,为协同推理阶段提供高质量的非结构化证据支持。根据综合得分 $\text{score_fn}(s_i)$ 对所有段落进行降序排序。然后按照预设的相似度阈值 η 和最小返回 $N_{\min}$ 进行筛选,即保留所有得分 $\text{score_fn}(s_i) \geq \eta$ 的段落。

1.5 闭环知识更新策略

为促使知识图谱从静态存储向动态持续演进系统转变,设计一种以多源证据验证为驱动的闭环知识更新机制。

1.5.1 更新触发与异常感知机制

不同于直接依赖LLMs主观判断知识是否过时,本文通过分析推理过程中的客观表现,引入3类异常信号作为知识更新的触发条件。1)空结果或弱关联子图。若图谱查询返回空结果,或检索子图的相关性得分低于设定阈值,则视为知识缺失,触发外部检索以补充缺失的事实。2)多跳推理链断裂。在处理多跳问题时,若中间子问题无法在知识图谱中链接到有效实体(如“下属舰船”无法映射到具体舰艇),则判定图谱结构不完整,需通过检索增强生成获取缺失的中间实体或关系。3)时效敏感问题与静态答案冲突。当用户问题包含明确时效指示词(如“现任”“最新”),而图谱返回的事实为静态或历史信息时,则触发对该信息当前有效性的验证。

1.5.2 多源证据验证与一致性判定

触发更新后,系统从权威外部源检索相关文档片段。大语言模型负责对图谱原有知识与外部证据进行融合与一致性判定。若验证结果显示二者在语义层面保持一致,将据此增强当前事实的置信度;若外部文档包含原图谱未覆盖的实体或关系,则识别为知识缺失并准备增补;若两者出现事实冲突且外部信息来源权威且多源一致,则判定原有图谱中的信息已过时,需要进行更新或替换。

1.5.3 知识结构化与图谱演进

通过验证的高置信度新知识将被结构化为标准三元组添加回知识图谱中,并支持附加时间属性。例如,确认“USS Carl Vinson的母港为圣迭戈海军基地”后,系统生成三元组(USS Carl Vinson, has_homeport, San Diego Naval Base),并可标注有效时间(如valid_from: 2025-Q1)。同时,更新记录(推理路径、证据来源、时间戳、置信度等)将作为元数据保存,形成可审计的知识演化轨迹。

通过这个闭环,知识图谱不再是一个静态的数据库,而是一个能够与外界交互不断学习和成长的动态知识系统,从而持续提升问答系统的知识储备和时效性。同时,闭环更新机制将知识更新从依赖主观判断转为任务驱动与证据验证相结合的过程,提升更新准确性与可解释性,适用于军事、医疗、金融等高时效性与高可靠性要求场景。

2 实验与分析

2.1 实验设置

2.1.1 数据集

本研究在WebQSP(WebQuestionsSP)和CWQ(Complex Web Questions)两个经典知识问答数据集上进行评估。两者均基于Freebase知识图谱构建,并提供自然语言问题及其对应的标准SPARQL查询。数据集统计信息如表1所示,外部知识源采用Wikidata作为补充文档数据库。

表1 实验数据集统计情况

数据集	问题数量	实体数量	关系数量	训练集数量	验证集数量	测试集数量
WebQSP	4 737	2 461	628	3 098	250	1 639
CWQ	34 689	11 422	845	27 639	3 519	3 531

2.1.2 评价指标

实验采用知识推理任务中广泛使用的准确率(accuracy)、命中率(Hit@k)与F1值作为评价指标。准确率与Hit@k反映模型在实体链接与路径预测上的准确性;F1值综合衡量分类任务的精确率与召回率,适用于整体性能评估。

2.1.3 对照模型

为比较模型性能,选取以下3类主流方法作为基线模型。信息检索类基准模型包括Subgraph Retrieval^[1]、Topic Unit^[15]、QGG^[7]和UniKGQA^[13];语义解析类基准模型包括Program Transfer^[16]、DecAF^[12]和HGNet^[17]模型。此外,为对比当前基于大语言模

型的知识推理方法,选取 ToG^[3]、ChatKBQA^[5]和 ToG-2^[8]作为所提模型的代表性对照模型。

2.1.4 参数设置

采用网格搜索法对模型的超参数进行选取,训练回合数(epoch)∈{10,50,100},训练批次(batch size)∈{1,2,3,4},学习率∈{5×10⁻⁵,5×10⁻⁴,5×10⁻³}。通过实验,本文所得的最优超参数设置为:训练回合数设置为100,训练批次设置为4,学习率设置为5×10⁻⁵,束搜索的大小设置为8。为使权重参数在数据集上取得最佳性能,在验证集上采用网格搜索方法进行调优,当权重组合取α=0.6、β=0.25、γ=0.15时,系统在复杂问题上的整体问答准确率最高,所以采用该设置。

文本编码使用 gte-large-en-v1.5 模型(434×10⁶参数),该模型在大规模文本嵌入基准(Massive Text Embedding Benchmark, MTEB)中表现均衡,兼顾效率与效果。实验在两张 NVIDIA A800 80 GB 的 GPU 上进行,结果取 5 次随机运行的平均值。由于查询缺乏真实的子图标注,本文沿用基于路径的弱监督方法,以主题实体与答案实体间的最短路径作为训练信号。

各环节的大语言模型调用配置(单次查询)具体如表 2 所示。大语言模型的分解与三元组抽取提示词示例如图 2 和图 3 所示。

表 2 大语言模型的具体配置及次数			
环节	调用次数	模型版本	提示策略
问题分解	1	Qwen3-32B	少样本思维链(3-shot)
融合推理与答案生成	1	Qwen3-32B	零样本+上下文注入
三元组抽取	1(仅更新触发时)	Qwen3-32B	少样本抽取模板(2-shot)

Please decompose the following complex question into a series of simple sub-questions, and specify the dependencies between sub-questions.
【Example】
Question: What movies did Ang Lee, who won the Best Director Oscar, direct
Decomposition:
Sub-question 1: Who won the Best Director Oscar → [Answer: Ang Lee]
Sub-question 2: What movies did Ang Lee direct → [Dependency: result of Sub-question 1]
【Task】
Question: {Q}
Please output in the following format:
Sub-question list: [sub-question1, sub-question2, ...]
Dependencies: [sub-question i → sub-question j] (meaning the result of sub-question i is the input for sub-question j)

图 2 分解提示词示例

Please extract all factual triples (head entity, relation, tail entity) from the following answer.
【Example】Answer: Ang Lee directed “Life of Pi” and “Brokeback Mountain”. Triples: (Ang Lee, directed, Life of Pi)(Ang Lee, directed, Brokeback Mountain)
【Answer】{generated_answer}

图 3 三元组抽取提示词示例

2.2 实验结果分析

SKG-RAG 与现有代表性方法在不同实验数据集上的问答准确率如表 3 所示,其中空白表示无此项。SKG-RAG 在 WebQSP 和 CWQ 两个数据集上均取得最优性能。在 WebQSP 数据集上,SKG-RAG 的准确率、Hit@1 和 F1 分别达到 81.6%、88.3% 和 85.2%,全面优于 ChatKBQA(73.8%、83.2%、79.8%),分别提升 10.6%、6.1% 和 6.8%。在更具挑战的复杂问答数据集 CWQ 上,SKG-RAG 的准确率、Hit@1 和 F1 分别为 79.8%、88.2% 和 83.4%。相比 ChatKBQA 的 73.3%、82.7% 和 77.8%,分别提升 8.9%、6.7% 和 7.2%,表现出良好的泛化能力与稳定性。

表 3 在不同数据集上的实验结果对比							单位:%
类别	模型	WebQSP			CWQ		
		准确率	Hit@1	F1	准确率	Hit@1	F1
信息检索类	Subgraph Retrieval		69.5	64.1		50.2	41.7
	Topic Unit		68.2	67.9		39.3	36.5
	QGG		73.0	74.0		44.1	40.4
	UniKGQA		77.2	72.2		51.2	49.4
语义解析类	Program Transfer		74.6	76.5		58.1	58.7
	DecAF		82.1	78.8		70.4	
	HGNet	70.7	76.9	76.6	57.8	68.9	68.5
大语言模型增强类	ToG		82.6	82.6		69.5	
	ChatKBQA	73.8	83.2	79.8	73.3	82.7	77.8
	ToG-2		84.1			81.6	
	SKG-RAG	81.6	88.3	85.2	79.8	88.2	83.4

传统基于信息检索或语义解析的方法(如 DecAF)在复杂多跳场景下性能受限,而仅依赖大语言模型的方法(如 ToG、ChatKBQA)虽具备较强的语义理解能力,却在事实一致性与知识更新方面存在不足。SKG-RAG 通过动态子图检索机制增强对结构化知识的精准利用,结合非结构化文本的混合检索以覆盖最新知识,并借助闭环知识更新策略实现系统的持续演进,从而在多项指标上显著领先,验证了本文协同推理框架的有效性。

2.3 消融实验

为验证 SKG-RAG 框架中不同模块的有效性,在 WebQSP 数据集上进行消融实验,结果如表 4 所示。实验中的 5 个模型,分别是-Graph-Doc(移除图检索与文档检索模块,仅依赖大语言模型进行推理)、-Doc(仅保留知识图谱检索模块,具体采用基于表示学习的知识图谱推理技术^[18])、-Graph(仅使用非结构化文档检索)、-Update(保留图检索与文档检索模块,但移除闭环知识更新机制,即不将新知识增量更新至知识图谱)以及 SKG-RAG 模型。SKG-RAG 融合图检索与文档检索,性能达到最优,准确率为 81.6%,*F1* 值为 85.2%,显著超越各消融版本。这表明结构化与非结构化知识在复杂推理任务中具有互补性;动态混合检索机制能够有效结合知识的准确性与覆盖面;多源知识协同增强了大语言模型的推理鲁棒性与事实一致性。

表 4 在 WebQSP 数据集上的消融实验结果对比 单位: %

模型	准确率	Hit@1	<i>F1</i>
-Graph-Doc	70.2	78.5	74.4
-Doc	74.8	83.1	79.3
-Graph	67.5	76.8	71.2
-Update	77.3	85.2	81.5
SKG-RAG	81.6	88.3	85.2

3 结束语

针对大语言模型知识静态、易产生幻觉与知识图谱覆盖不全、更新滞后的挑战,提出了一种面向动态知识检索的大语言模型与知识图谱协同推理框架 SKG-RAG。该框架通过结构感知的子图检索与多维度文档匹配,实现了多源知识的精准获取;借助闭环更新机制,可将推理过程中产生的高置信度新知识持续反馈至知识图谱,实现系统的自我演化。实验表明,该方法能有效处理复杂多

跳问题,提升答案的准确性与可溯源性,并保障知识库的时效性。

未来工作将重点研究:建立更高效的知识验证机制,提升知识更新质量;扩展至多模态场景,构建融合文本、图像等的多模态知识推理系统;探索基于动态知识图谱的智能体系统,增强长期记忆与自主学习能力。

参考文献:

[1] Zhang J, Zhang X K, Yu J F, et al. Subgraph retrieval enhanced model for multi-hop knowledge base question answering [C]//Proceedings of the 60th annual meeting of the association for computational linguistics. Stroudsburg, USA: ACL, 2022: 5773-5784.

[2] Jiang J H, Zhou K, Dong Z C, et al. StructGPT: a general framework for large language model to reason over structured data [C]//Proceedings of the 2023 conference on empirical methods in natural language processing. Stroudsburg, USA: ACL, 2023: 9237-9251.

[3] Sun J, Xu C, Tang L, et al. Think-on-graph: deep and responsible reasoning of large language model on knowledge graph [C]//Proceedings of the 12th international conference on learning representations. Appleton, USA: ICLR, 2024: 3868-3898.

[4] Shu Y H, Yu Z W, Li Y H, et al. TIARA: multi-grained retrieval for robust question answering over large knowledge base [C]//Proceedings of the 2022 conference on empirical methods in natural language processing. Stroudsburg, USA: ACL, 2022: 8108-8121.

[5] Luo H R, E H, Tang Z C, et al. ChatKBQA: a generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models [C]//Findings of the association for computational linguistics: ACL 2024. Stroudsburg, USA: ACL, 2024: 2039-2056.

[6] Bhutani N, Zheng X Y, Jagadish H V. Learning to answer complex questions over knowledge bases with query composition [C]//Proceedings of the 28th ACM international conference on information and knowledge management. New York, USA: ACM, 2019: 739-748.

[7] Lan Y S, Jiang J. Query graph generation for answering multi-hop complex questions from knowledge bases [C]//Proceedings of the 58th annual meeting of the association for computational linguistics. Stroudsburg, USA: ACL, 2020: 969-974.

[8] Ma S J, Xu C J, Jiang X H, et al. Think-on-graph 2.0: deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation [C]//Proceedings of the 13th international conference on learning representations. Appleton, USA: ICLR, 2025: 52782-52806.