天津科技大学学报
Journal of Tianjin University of Science & Technology
ISSN 1672-6510,CN 12-1355/N

《天津科技大学学报》网络首发论文

题目：基于多结构图检索的大语言模型增强方法
作者：王嫄，宾茂杰，李航举，刘文悦，赵婷婷，吴骏
DOI：10.13364/j.issn.1672-6510.20250153
收稿日期：2025-10-11
网络首发日期：2026-09-02
引用格式：王嫄，宾茂杰，李航举，刘文悦，赵婷婷，吴骏．基于多结构图检索的大语言模型增强方法[J/OL]．天津科技大学学报．
https://doi.org/10.13364/j.issn.1672-6510.20250153

DOI：10.13364/j.issn.1672-6510.20250153

基于多结构图检索的大语言模型增强方法

王　媛[1]，宾茂杰[1]，李航举[1]，刘文悦[1]，赵婷婷[1]，吴　骛[2]
(1. 天津科技大学人工智能学院，天津 300457；2. 四川大学华西医院，成都 610041)

摘　要：大语言模型（LLM）在实际应用场景中面临知识时效性局限、幻觉生成和决策黑盒三大挑战，现有基于知识图谱的增强方法存在图结构丢失、检索焦点过窄和信息冗余方面的缺陷。为此，本文提出多结构图检索增强方法（multi-structure graph retrieval augmented method, MSGR），设计整合精确子图匹配与支链推理扩展的双通道检索机制，精确子图匹配基于实体相似度构建局部连通结构，保留关键语义以及精确图结构信息；支链推理扩展通过多跳结构探索拓宽检索焦点，利用大语言模型筛选和捕获最小冗余隐性关联知识。通过加权集成与关系去重精炼上下文信息，融合生成可靠的结构化上下文。在罕见病诊断任务测试中，MSGR 在 RAMEDIS 数据集上诊断准确率达到了 0.795，相较链式检索 ToG 提升了 7.5%，相较子图检索 MindMap 提升了 17.5%，所提方法能有效缓解知识缺失与幻觉问题，MSGR 在 LIRICAL、PUMCH 等数据集的所有指标均优于最新同类型方法（如 ToG、MindMap、KG^{2}RAG），本文方法已在医疗场景验证，未来将扩展至多领域。

关键词：大语言模型；知识图谱；检索增强；多结构图检索
中图分类号：TP391.1　　**文献标志码：**A

Large Language Model Enhancement Method Based on Multi-Structure Graph Retrieval

WANG Yuan[1]，BIN Maojie[1]，LI Hangju[1]，LIU Wenyue[1]，ZHAO Tingting[1]，WU Qin[2]
(1.College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China;
2. West China Hospital, Sichuan University, Chengdu 610041, China)

Abstract: Large language models (LLMs) face three major challenges in practical applications: limitations in knowledge timeliness, hallucination generation, and decision black boxes. Existing knowledge graph-based enhancement methods suffer from drawbacks such as loss of graph structure, overly narrow retrieval focus, and information redundancy. To address these issues, this paper proposes a Multi-Structure Graph Retrieval Augmented method (MSGR), which designs a dual-channel retrieval mechanism integrating precise subgraph matching and branch inference expansion. Precise subgraph matching constructs locally connected structures based on entity similarity, preserving key semantics and accurate graph structure information. Branch inference expansion broadens the retrieval focus through multi-hop structural exploration, leveraging large language models to screen and capture minimally redundant latent associative knowledge. Contextual information is refined via weighted integration and relationship deduplication, fusing to generate reliable structured context. In rare disease diagnosis benchmark tests, MSGR achieved a diagnostic accuracy of 0.795 on the RAMEDIS dataset, representing a 7.5% improvement over the chain-based retrieval method ToG and a 17.5% improvement over the subgraph-based method

收稿日期：2025–10–11；**修回日期：**2026–03–19
基金项目：国家重点研发计划项目（2022YFC2504500）；四川省自然科学基金项目（2024NSFSC1532）；天津市自然科学基金资助项目（24JCYBJC00890）
作者简介：王　媛（1989—），女，山西万荣人，副教授；**通信作者：**赵婷婷，教授，tingting@tust.edu.cn

MindMap. The proposed method effectively mitigates knowledge gaps and hallucination issues. MSGR outperforms the latest comparable methods (e.g., ToG, MindMap, KG2RAG) across all metrics on datasets such as LIRICAL and PUMCH. While validated in medical scenarios, the method is designed for future extension to multiple domains.

Key words: large language models; knowledge graph; retrieval-augmented generation; multi-structure graph retrieval

近年来，GPT 系列[1-2]、LLaMA[3]、PaLM[4]、DeepSeek[5]等大语言模型的迭代演进显著提升了文本生成、逻辑推理及跨语言迁移等任务的性能边界。但在实际应用中仍面临三重核心挑战：第一，知识时效性：大语言模型知识更新不及时，通过预训练或微调还可能会导致模型产生知识遗忘[6]；第二，幻觉生成：模型易生成语义合理但事实错误的输出[7]，在医疗诊断等高风险场景存在严重隐患；第三，决策黑盒：黑盒特性导致推理过程不可追溯[8]，知识隐式存储于参数中难以验证，神经网络推理机制亦缺乏可解释性。

针对上述挑战，引入知识图谱（knowledge graph，KG）等外部知识源已成为提升大语言模型生成质量的一种重要技术路径。此类结构化知识的引入能够显著增强大语言模型输出内容的准确性与完整性[9]。大语言模型自身具备的上下文学习能力为外部知识的融合提供了基础，通过在输入上下文中嵌入与任务相关的背景信息，能够有效纠正模型生成过程中的事实性错误，并提升其对未知或动态知识的理解与适应能力。

基于知识图谱增强大语言模型的主要通过从结构化知识源中检索与查询相关的信息，并将其以自然语言或结构化形式注入大语言模型的上下文窗口中，以弥补模型内部知识在时效性、准确性和可解释性方面的不足[10-11]。该增强流程通常包含 3 个关键环节：首先，从知识图谱中执行检索，获取与用户查询相关的实体、关系或子结构；其次，对检索结果进行筛选、去重与表示转换，形成适合模型理解的输入格式；最终，将处理后的知识信息与大语言模型集成，辅助其生成更可靠、更准确的回答[12-14]。

现有的知识图谱增强大语言模型方法存在局限性。基于三元组或实体节点的检索方法会出现图结构的丢失，导致检索的信息缺乏逻辑性与连贯性[15]。这类方法通常从知识图谱中孤立地抽取与查询相关的三元组或实体节点，或者将三元组或实体节点简单罗列作为大语言模型输入。检索过程破坏了知识图谱固有的、丰富的网状连接关系。实体之间的多跳关系、路径依赖和全局语义关联在检索后被剥离，仅保留了关联信息不足、离散的知识碎片。基于推理链的检索方法[16]会导致检索结果仅集中于相关性最高的单条信息链，从而忽略次要信息中可能蕴含的补充知识。这类方法旨在模拟人类的逐步推理过程，通过构建概率最高或最显著的单一线性的、高相关性的推理链回答复杂问题。该方式为了追求推理的效率和高置信度，在许多场景下，问题的解答往往需要综合多条并行或次要的线索，而非依赖单一的推理路径。基于子图的检索方法[17]会出现检索信息过于繁杂冗余，重点信息不突出从而导致模型无法关注到对于回答用户问题最重要的信息，最后出现回答重点偏离的情况。基于子图的检索方法检索结果是信息非常丰富的子图，该方式虽然附带的信息丰富，但是无法保证这些信息的实用性，往往大多信息是无用的，这就会导致模型注意力分散在众多无用信息中。基于全局图结构的社区检索方法虽然能够通过社区发现算法捕获知识图谱中的远距离语义关联，但存在社区划分静态化与计算复杂度高的双重挑战。这类方法（如 GraphRAG[18], HiRAG[19]）依赖于预定义的社区结构，难以适应不同查询的个性化语义需求：社区划分的粒度固定导致检索范围缺乏弹性，可能产生信息过宽（引入无关噪声）或过窄（遗漏关键证据）的问题；同时，大规模图谱的社区发现与维护需要高昂的计算开销，限制了其在实时性要求较高场景中的应用可行性，且在动态更新的知识图谱中面临社区结构频繁重构的挑战。

针对现有基于知识图谱的增强方法普遍存在的图结构丢失、检索焦点过窄与信息冗余三大核心问题，本文创新性地提出了多结构图检索增强方法（multi-structure graph retrieval augmented method, MSGR）。MSGR 是一种深度融合、结构互补的双通道检索机制，其核心思想是通过异构图结构的协同与互补，实现检索效果在准确性、覆盖度与可解释性三个维度上的共同提升。一方面通过精确子图匹配结构检索，在局部范围内构建语义连贯、结构完整的子图，有效缓解了传统三元组检索中的结构破碎与语义割

裂问题，确保核心知识的完整性与上下文关联性；另一方面，通过支链推理扩展结构检索突破单一子图的范围限制，依托大语言模型的推理能力对潜在多跳关系进行智能探索与路径筛选，显著扩展检索广度与隐式关联挖掘能力，从而避免检索焦点过窄所带来的信息缺失。

1　相关工作

基于知识图谱的大语言模型检索增强方法是一种通过结构化知识表示（以实体为节点，以关系为边）增强大语言模型推理能力的技术。该方法首先从非结构化文本中提取实体、关系及属性，构建具有语义关联的知识图谱；随后通过图嵌入技术将节点与关系向量化，建立可检索的索引库。查询时，系统识别查询中的实体，遍历知识图谱获取关联子图（包括多跳关系），并利用加权融合策略（如基于余弦相似度的权重分配）整合检索结果。最终，将结构化的图谱信息还原为自然语言上下文，输入生成模型以产生精准且具备逻辑连贯性的答案。此法显著提升了复杂查询（如多跳推理、跨域知识整合）的处理能力，并有效缓解传统检索增强在语义歧义性和幻觉问题上的局限。目前主流的方法有 4 种，包括基于三元组与实体检索、推理链检索、子图检索、全局图结构社区检索。

1.1　三元组与实体检索

基于三元组与实体检索方法是一种通过从知识图谱中提取离散的结构化知识单元（实体及其关系），为大语言模型提供外部知识支持的技术范式。这类方法的核心在于将知识图谱分解为原子化的三元组（头实体-关系-尾实体）或实体节点，通过向量化表示与相似度计算，检索与查询最相关的结构化知识片段，并将其注入大语言模型的上下文窗口，以增强其生成质量。Kang 等[20]在知识图谱中提取出问题相关实体三元组，即将主语、关系和宾语文本拼接起来后，输入问题提示信息中，以增强模型能力。

这种基于三元组与实体的检索方法通常存在图结构丢失、推理深度不足的问题，对于需要多步推理的复杂问题，这类方法往往得不到正确结果。上述问题限制了现有方法在实际应用中的实用性，而本文提出的多结构图检索方法包括多种不同的推理图结构，通过精确子图检索，根据用户问题关键词精确检索构建子图，解决了图结构丢失问题。通过支链推理扩展对检索的实体进行基于相关性的多步推理扩展，解决了上述的推理深度不足问题。

1.2　推理链检索

基于关键推理链的知识图谱增强方法，是一种通过从知识图谱中识别并提取与当前查询最相关的一条高置信度推理路径（通常为多跳关系链），将其转化为连贯的自然语言表述，并作为增强上下文注入大语言模型的技术。这类方法的核心在于利用图推理算法（如基于关系权重或语义相似度的路径搜索）从知识图谱中的问题实体出发，遍历多跳关系，筛选出得分最高的单条路径作为关键推理链。此推理链不仅提供了显式的结构化知识，还通过序列化的逻辑步骤（如“实体 A→关系→实体 B→关系→实体 C”）为大语言模型构建可解释的推理上下文，从而显著提升模型在复杂问答（尤其是多跳推理）中的事实准确性与逻辑连贯性。相较于基于子图或全邻域检索的方法，此类方法通过聚焦单条高相关路径，有效降低了上下文冗余与计算开销。Zhang 等[21]通过设置特殊规则，结合知识图谱中多个三元组的信息，生成知识链，使用提示词工程设置 prompts，将知识链中信息整合到输入提示中。Jiang 等[22]设计预检索阶段，通过大语言模型返回假设相关实体结合问题实体进行推理链检索，扩展信息广度。Ma 等[16]提出在图谱上思考，在推理链的检索中与大语言模型进行深度交互，由大语言模型决定检索内容质量，从而检索出高质量推理链。

然而上述基于推理链的检索方法仅关注基于单个实体的深度检索，虽然对于单个实体能够检索出更多信息，但是基于推理链的方法也存在图结构缺失的问题，并且仅仅基于单条推理链无法保证检索信息的丰富程度，这种情况在医疗、金融等需要大量相关参考信息的领域尤为突出。在本文 MSGR 框架中，通过并行添加基于用户查询中关键实体的精确子图匹配，给检索结果加入了知识图谱中更多相关信息补充结果的信息丰富程度，保证检索结果内容能够更好地回答用户问题。

1.3　子图检索

基于推理子图检索的知识图谱增强方法是一种通过从大规模知识图谱中识别并提取与查询语义相关且富含多跳推理路径的连通子图，将其转化为结构化表并注入大语言模型上下文的技术。这种方法首先从查询中识别核心实体，使用余弦相似度在向量数据库中检索与关键词最相关的实体，筛选构建出覆盖潜在推理路径的紧凑子图；随后将子图以自然语言形式

输入大语言模型，并利用思维链机制引导模型基于子图证据进行多步推理。相较于原子化三元组检索，这种方法通过保留局部图结构（如邻域关系与路径依赖），能够显著提升复杂问答（尤其是多跳推理）的准确性、可解释性及幻觉抑制能力。Edge 等[18]通过查询中的核心实体，从知识图谱中提取这些实体周围的多跳邻居，形成一个丰富的上下文子图作为大语言模型的上下文输入。Wen 等[17]通过识别问题中的实体，并从知识图谱中检索这些实体的多跳路径和直接邻居形成证据子图，为后续推理提供结构化信息。

上述基于检索子图的检索方法存在信息冗余与单子图检索，仅能保证检索结果的广度，而无法保证基于查询实体的检索深度，会导致针对复杂问题的检索深度不足，从而无法获得正确信息。在本文的框架中，通过支链推理扩展基于用户查询实体进行多步推理检索出一条推理链，扩展用户查询实体的检索深度，以保证能够精确回答需要多步推理的复杂问题。

1.4　全局图结构社区检索

全局图结构社区检索是一种基于社区发现算法，将知识图谱划分为语义稠密子图社区，通过检索与查询相关的完整社区捕获全局语义关联的创新方法，能够有效捕获实体间的远距离语义关联和隐含的群体知识模式，显著提升了复杂问答任务中的知识覆盖广度，但同时也因社区结构的静态性和计算开销较大。

GraphRAG[18]作为基于全局图结构检索的代表性方法，通过社区发现算法（如 Louvain 算法）将知识图谱划分为语义社区，检索时首先定位查询相关的社区，再从社区中提取关键子图作为增强上下文。这种方法能够捕获远距离的语义关联，特别适合需要全局知识理解的复杂问答任务。

然而，GraphRAG 这类基于社区检索的方法也存在明显局限性。首先，社区划分的粒度难以自适应调整，可能导致检索范围过宽（引入噪声）或过窄（遗漏关键信息）；其次，静态的社区结构难以动态响应查询的具体语义需求；此外，大规模图谱的社区计算开销较高，影响实时性。相比之下，HiRAG 方法[19]通过层次化的社区划分策略，在保持全局语义关联的同时，实现了社区粒度的动态调整，但其仍然面临社区结构预定义的限制，在实时性要求较高的场景中应用受限。

本文 MSGR 方法属于图检索增强范式下的创新探索，通过双通道机制整合局部子图匹配与动态推理链扩展的优势。与 GraphRAG 的静态社区检索不同，MSGR 采用动态路径发现与语义筛选策略，既保持了结构完整性，又实现了检索范围的精准控制，为复杂场景下的知识增强提供了更灵活的解决方案。

2　算法设计

2.1　算法整体框架

本文算法框架如图 1 所示。算法整体流程表示为

$$C^{*}=F_{\text{fusion}}\left(S_{\text{subgraph}}\left(q,G\right),C_{\text{chains}}\left(q,G\right)\right) \tag{1}$$

其中：S_{subgraph} 为精确子图匹配函数，基于用户查询 q 和知识图谱 G 中的实体相似度，构建局部连通子图，确保结构完整性和语义连贯性；C_{chains} 为支链推理扩展函数，通过多跳关系探索和大语言模型筛选捕获隐性关联知识，扩展检索广度；多结构融合函数 F_{fusion} 则对双通道输出进行加权集成与关系去重，生成最优上下文 C^{*}。

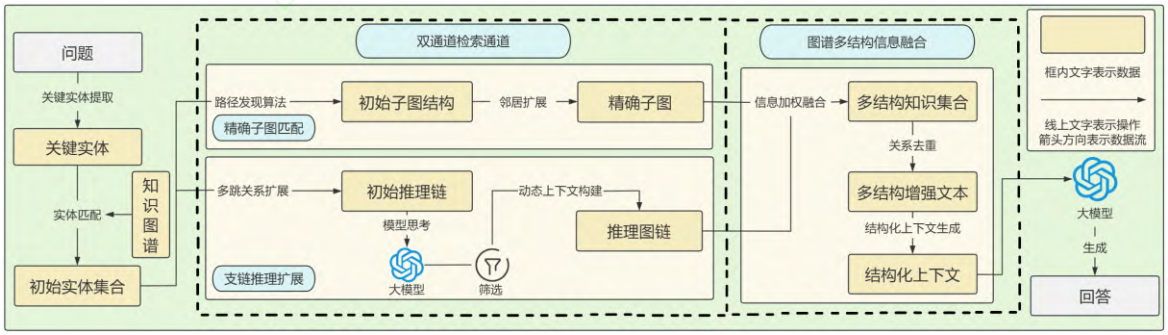


图 1　MSGR 整体框架图

Fig. 1　Overall framework of MSGR

算法首先通过实体匹配，结合提示词工程使用大语言模型从用户问题 q 中提取关键实体，然后通过双通道检索模块 S_{subgraph} 与 C_{chains} 并行获取两种异构结构：精确子图匹配结构经由实体向量相似度匹配，从

用户查询中提取关键实体并初始化集合，通过路径发现算法构建实体间初始子图结构，并利用邻居扩展机制增强上下文完整性；支链推理扩展结构则通过多跳关系扩展，从初始实体集合出发探索潜在推理路径，借助大语言模型进行语义筛选与动态上下文构建；获取上述结构后，图谱多结构信息融合模块 F_{fusion} 对双通道输出进行加权集成，执行基于实体对标准化的关系去重，最终转化为 CSV 格式的结构化上下文，以供大语言模型生成可靠回答。

本文将知识图谱中的实体关系通过向量模型 Bge-m3，全部转换成向量形式，存储为实体关系向量数据库，以便后续进行相似度匹配。

2.2　实体匹配

通过大语言模型从用户问题中提取出关键实体 e_i，将其转化为关键实体向量 $\vec{k}$，为了针对用户问题中的关键实体 e_i 进行范围检索，检索出初始实体集合 E_{initial}。设计精确子图匹配流程，通过向量相似度进行实体匹配，匹配出初始实体集合 E_{initial}。设实体向量数据库为 D_{e}，用于存储知识图谱中所有实体的向量表示，实体向量数据库中对应的关键实体向量表示为 $\vec{e}$。实体匹配过程定义为

$$E_{initial} = e_i \mid \mathrm{sim}\left(\vec{k}, \vec{e_i}\right) > \tau, e_i \in D_{\text{e}} \tag{2}$$

其中：sim( )为余弦相似度函数，τ 为相似度阈值。

2.3　精确子图匹配结构构建

2.3.1　路径发现算法

针对在实体匹配中检索出来的实体 e_i 进行路径探索，将游离零散的实体链接成初始子图结构 P_{subgraph}。

$$P_{subgraph} = \bigcup_{e_i, e_j \in E_{initial}, \mathrm{i} \neq \mathrm{j}} FindPath\left(e_i, e_j, G\right) \tag{3}$$

对于初始实体集合 E_{initial}，构建实体间的关系路径，为

$$FindPath\left(e_i, e_j, G\right) = \bigcup_{p \in Paths\left(e_i, e_j, G\right)} \left(p, w\right) \tag{4}$$

其中：p 为路径集合 $Paths(e_i,e_j,G)$中的一条路径，w 为路径集合总权重，p 与 w 由如下公式得出。

$$p = \left(R^1, w^1\right), \left(R^2, w^2\right), \ldots, \left(R^m, w^m\right), \left(R^{m+1}, w^{m+1}\right) \tag{5}$$

其中：R^m 是连接第 m 个与第 m+1 个实体的关系。

$$w = \Sigma_{m=1}^{n} w^m \, \delta_m \tag{6}$$

其中：n 为检索路径中的边数；w^m 为第 m 条边的权重，这些权重来源于知识图谱中每条边的置信度属性；δ_m 为路径衰减因子，δ_m 越接近 1，算法会更倾向于探索和保留更长的路径，检索范围更广，δ_m 越接近 0，算法会更加聚焦于短路径，强调检索结果的精确性和紧密度。δ_m 的取值通过网格搜索方法确定，搜索范围为{0.1-0.9}，以 0.1 为步长。调优过程基于验证集 PUMCH 上 hit@1 指标的综合性能评估，最终选择 δ_m=0.6 作为最优值，以保证路径长度与检索精度间取得最佳平衡。

2.3.2　邻居扩展机制

为增强上下文完整性以提升检索内容质量，对高相关度实体进行邻居扩展，为

$$N_{expand} = \bigcup_{e \in TopK\left(E_{initial}, k\right)} Neighbors\left(e, G\right) \tag{7}$$

其中：N_{expand} 表示扩展后得到的邻居实体集合，e 依次代表集合 TopK($E_{initial}$,k) 中的每一个实体。Neighbors(e,G)表示从知识图谱 G 中，查询并返回与实体 e 直接相连的所有邻居实体。

邻居扩展权重计算公式为

$$w_{neighbor}\left(e_n, e_c\right) = \alpha \cdot sim\left(e_n, q\right) + \beta \cdot \deg ree\left(e_n\right) + \gamma \cdot edge_{weight}\left(e_c, e_n\right) \tag{8}$$

其中：α、β、γ 为权重系数，α 为语义相似度权重，β 为节点度数权重，γ 为边权重；e_c 为中心实体，$e_c \in E_{initial}$；e_n 为邻居实体；$w_{neighbor}$ 为邻居实体扩展的综合权重得分，$degree(e_n)$为实体 e_n 的度数（连接边数），$edge_{weight}(e_c, e_n)$为实体对之间的边权重。邻居扩展权重部分通过加权求和计算邻居实体 e_n 的扩展权重。

2.4　支链推理扩展结构构建

2.4.1　多跳关系扩展

在实体匹配的基础上，算法通过多跳关系扩展机制进一步探索与初始实体集相关联的潜在推理路径。扩展过程中依据迭代公式进行实体集合的更新，即

$$E_{chains}^{(t+1)} = \bigcup_{e \in E_{chains}^{(t)}} Explore\left(e, G, w_{search}\right) \tag{9}$$

其中：$E_{chains}^{(t+1)}$ 表示在第 t+1 次迭代后，得到的新实体推理链集合；w_{search} 为搜索宽度参数，控制每一跳扩展的实体数量；探索函数 $Explore(e,G,w)$定义为从实体 e 出发，根据知识图谱 G 中的关系边，选择 $TopK$ 个相邻实体进行扩展得到多条初始推理链，其形式化表示为

$$Explore\left(e, G, w\right) = TopK\left(e' \mid \left(e, r, e'\right) \in E \text{ or } \left(e', r, e\right) \in E, w\right) \tag{10}$$

其中：e' 为候选邻居实体，r 为连接实体 e 与 e' 的关系，E 为知识图谱中的边集合。

2.4.2　基于大语言模型的筛选

为控制路径搜索空间并提升检索质量，算法引入基于大语言模型的筛选机制，对候选关系进行语义层

面的筛选与排序。筛选过程表示为

$$R_{selected} = LLM_{Select}\left(R_{candidates}, q, w_{search}\right) \quad (11)$$

其中：$LLM_{\text{Select}}()$表示使用大语言模型对候选关系 $R_{\text{candidates}}$ 进行筛选排序，筛选出相关性高的关系 R_{selected}。

2.4.3 动态上下文构建

为了保证上下文在推理过程中根据实际情况进行实时变更，使用动态上下文构建方法，在支链推理过程中动态维护上下文状态，公式为

$$C_{chains}^{(t)} = C_{chains}^{(t-1)} \cup ContextUpdate\left(E_{new}^{(t)}, R_{new}^{(t)}\right) \quad (12)$$

其中：$ContextUpdate(\cdot,\cdot)$表示将新扩展出来的实体 $E_{new}^{(t)}$ 与关系 $R_{new}^{(t)}$ 更新到推理链中的上下文更新函数，$C_{chains}^{(t)}$ 在第 t 步推理迭代后，最终生成并维护的动态上下文知识库。通过迭代更新，能够灵活地捕捉复杂查询中所需的多样化知识片段，增强最终生成答案的准确性与鲁棒性。

2.5 图谱多结构信息融合

2.5.1 信息加权融合

本模块采用加权融合策略整合子图匹配和支链推理的结果，为

$$C_{fused} = M\left(C_{subgraph}, C_{chains}\right) \quad (13)$$

其中：C_{fused} 表示得到的多结构知识集合；C_{subgraph} 表示精确子图匹配结构；C_{chains} 表示支链推理扩展结构；融合函数 M 包含去重和权重平衡，M 包含的关系去重操作（$Dedup$）本质上是基于证据权重的冲突消解机制，定义为：

$$\begin{aligned} &M\left(C_1, C_2\right) = Dedup\left(C_1 \cup C_2\right) \\ &\text{with } w_{final}(x) = \sum\nolimits_{i} \omega_i \cdot w_i(x) \end{aligned} \quad (14)$$

其中：ω_i 为各结构的融合权重；$Dedup$ 为去重操作；ω_i 为各结构的融合权重，子图通道权重为 0.6，支链通道权重为 0.4；$w_i(x)$为知识单元 x（如实体、关系或路径）在检索通道 i 中的原始权重得分，反映了 x 在特定通道内的局部重要性；通道 i 包括两个分支：精确子图匹配通道（i=subgraph）和支链推理扩展通道（i=chains），每个通道根据其检索机制独立计算权重。通过动态权重分配机制，确保关键信息获得更高权重，同时平衡不同结构的贡献度。融合函数 M 先对双通道结果去重（$Dedup$），再通过权重 ω_i 平衡各结构贡献。权重 ω_i 基于网格搜索优化，确保关键信息主导。

2.5.2 关系去重

关系去重基于实体对标准化，表示为

$$Dedup(R) = r \mid Normalize(r.src, r.tgt) \notin Seen\_pairs \quad (15)$$

其中：R 为原始关系集合，包含所有待处理的关系，每个关系 r 包含源实体 $r.src$ 和目标实体 $r.tgt$；标准化步骤 $Normalize()$包括大小写统一、空格处理、方向无关化（A->B=B->A）、生成唯一键（创建标准化的实体对标识符）；$Seen\_pairs$ 表示已处理过的实体集合。如果标准化后的实体对不在已有集合中，则保留该关系；如果标准化后的实体对已存在于集合中，则丢弃该关系。

2.5.3 结构化上下文生成

最终上下文 C_{final} 采用 CSV 格式表示，为

$$C_{final} = \text{CSV}\left(E_{fused}\right) \oplus \text{CSV}\left(R_{fused}\right) \oplus \text{CSV}\left(T_{units}\right) \quad (16)$$

其中：$\oplus$表示拼接操作；$\text{CSV}(\cdot)$表示将文本转化为 CSV 格式；实体表 E_{fused} 包含编号 id、实体名称、实体类型、描述、来源；关系表 R_{fused} 包含编号 id、源实体、指向实体、描述、关键词、关系权重、来源；文本表 T_{units} 包含编号 id、文本内容。

通过拼接得到最终上下文 C_{final}，将 C_{final} 作为辅助信息，结合提示词输出给大语言模型，生成回答 Answers，为

$$Answers = LLM\left(prompt + C_{final}\right) \quad (17)$$

3 实验与结果分析

3.1 实验设计

3.1.1 语料库

本文使用的知识图谱是通过大语言模型对罕见病语料库进行知识抽取自动化构建的罕见病专项知识图谱。使用的语料库包含两部分：RareDis[23]和 ReCOP[24]。上述罕见病专项知识图谱包含 93114 个实体和 151800 个关系。

RareDis：由美国国家罕见疾病组织（NORD）建立和维护的罕见疾病数据库，包含 1041 个文本条目，涵盖 1200 种罕见疾病，附有罕见疾病及其临床表现的注释。

ReCOP：来源于国家罕见疾病组织（NORD）数据库，包含 1324 个文本条目，对应 1324 种罕见疾病。

3.1.2 数据集

HMS 数据集：该数据集来源于 Ronicke 等[25]发表的工作，包含 88 例临床病例，涉及 39 种罕见疾病。

LIRICAL 数据集：来源于 Zenodo 平台的罕见病

例，包含 370 例病例，涵盖 252 种罕见疾病。

MME 数据集：来源于 GA4GH MME API 测试数据集，包含 40 个病例，涵盖 17 种罕见疾病。

PUMCH 数据集：来源于 Chen 等[26]发表的工作，其中纳入了来自北京协和医院的真实临床病例，包括与 34 种罕见病相关的 75 例病例。

RAMEDIS 数据集：来源于 Töpel 等[27]的研究，包含 624 例病例，涉及 63 种罕见疾病。

3.1.3　评估指标

使用名为hit@n专门设计用于评估模型在诊断罕见疾病方面的性能。如果模型生成的前 n 个结果中至少存在一个正确诊断，则此度量被定义为成功。hit@1 度量，在提示中明确包含“仅输出一个诊断”指令；同样，对于 hit@n，允许模型输出前 n 个最可能的诊断。在本研究中，选择 hit@1、hit@3 以及 hit@5 作为综合评价模型诊断性能的主要评价指标。

3.1.4　对比方法

Naïve RAG：通过两个阶段建立基线 RAG 架构，文本分割/嵌入到向量数据库，然后通过最近邻搜索查询向量化和检索语义匹配的块，通过直接向量对齐确保计算效率。

CoT[28]：思维链方法通过连接输入查询、顺序推理步骤和最终输出的结构化示例，增强 LLM 中的推理。

LightRAG[29]：该方法抽象提炼跨文档的高层主题描述，处理查询时，提取查询中的全局关键词直接匹配高层主题，通过图谱遍历整合多源异构文本中的相关知识簇，突破传统分块检索的文档边界限制，实现无需实体定位的跨文档语义聚合。

Base：基于知识图谱的基础关键词三元组的检索方法，在知识图谱中检索用户问题中提取的关键词生成三元组作为外部知识。

ToG[16]：基于用户问题在知识图谱中检索关键推理链，通过与模型交互去除无用信息，构建精准且关键的逻辑推理链，将推理链作为知识输出给大语言模型。

MindMap[17]：从问题中识别关键实体，利用相似度匹配外部知识图谱节点，构建证据子图采用复合提示模板，引导 LLM 融合图谱证据与自身隐式知识，生成包含总结答案、推理链及可验证的思维导图的三段式输出。

KG^{2}RAG[30]：在离线阶段为文档库构建关联知识图谱，将非结构化文本转化为结构化的事实网络。先获取“种子”片段，然后通过图谱遍历扩展检索范围。利用图谱作为骨架，对检索到的杂乱片段进行过滤和重组织，生成结构清晰、内容连贯的叙事性段落后，再输入给大语言模型。

3.1.5　实验设置

采用 glm-4-flash 作为基座模型进行实验实现，该模型是由智谱 AI 发布的轻量化模型，参数量为 80 亿（8B），使用的是未经量化的版本。本文所使用的 GPU 配置为 3090（24 GB），如果要本地化部署模型，至少需要 16 GB 显存 GPU。如果考虑使用供应商提供的 API 接口进行部署，则对显存无要求。文本分段大小配置为 400 个 tokens，分段文本上下文窗口扩展设置为 100 个 tokens。采用的向量模型为 BAAI 发布的 Bge-m3。在整个实验过程中，使用 Nano 向量数据库对矢量数据进行存储和管理。

对于超参数的选择，取值依据均为 PUMCH 数据集中 hit@1 指标性能最优。相似度阈值 τ=0.75，取值范围为 0.5~0.9，路径衰减因子 δ_m=0.6，邻居扩展权重 α、β、γ 分别为 0.5、0.2、0.3，搜索宽度参数 w_{search} 为 4。权重 ω_i 为 0.6，该取值来源于图检索的广度与深度的权衡。在实际检索中，子图通道权重占比 0.6，以确保结构完整性优先；支链通道权重占比 0.4，以确保检索结果包含推理深度补充。

3.2　实验结果与分析

本文方法在 5 个数据集上性能结果见表 1。在除 MME 的 hit@1 指标上都达到了最优的性能，由此可知，本文提出的多结构图检索方法性能优于现有的主流知识图谱检索增强方法。在 RAMEDIS 数据集上诊断准确率达到了 0.795，相较链式检索 ToG 提升了 7.5%，相较子图检索 MindMap 提升了 17.5%。虽然在 MME 的 hit@1 指标上略低于 ToG，但是在后续的 hit@3 与 hit@5 中分别拉开了 12.5%与 7.5%的差距。在 HMS 数据集上相比于其他最优方法，在准确率上分别提升了 4.6%、1.1%、6.8%。在 LIRICAL 数据集上的性能在 hit@1、hit@3、hit@5 上分别提升 1.6%、0.8%、1.0%。在 PUMCH 数据集上的提升范围在 2.7%~8%。

表 1　使用相同基座模型与其他方法在罕见病诊断任务比较

Tab. 1　Comparison of the same base model and other methods in the task of rare disease diagnosis

方法	指标	诊断准确率				
		HMS	LIRICAL	MME	PUMCH	RAMEDIS
Naive	hit@1	0.261	0.227	0.300	0.333	0.364

RAG	hit@3	0.284	0.332	0.575	0.293	0.365
	hit@5	0.284	0.400	0.550	0.320	0.375
	hit@1	0.182	0.232	0.125	0.253	0.434
CoT	hit@3	0.364	0.405	0.475	0.333	0.625
	hit@5	0.409	0.405	0.350	0.253	0.577
	hit@1	0.239	0.119	0.225	0.360	0.202
Base	hit@3	0.114	0.122	0.225	0.267	0.258
	hit@5	0.114	0.108	0.200	0.280	0.274
	hit@1	0.239	0.389	**0.625**	0.387	0.585
ToG	hit@3	0.500	0.560	0.750	0.520	0.720
	hit@5	0.466	0.614	0.775	0.560	0.766
	hit@1	0.193	0.311	0.275	0.267	0.503
LightRAG	hit@3	0.386	0.527	0.625	0.387	0.705
	hit@5	0.477	0.530	0.575	0.453	0.768
	hit@1	0.284	0.373	0.400	0.360	0.590
MindMap	hit@3	0.284	0.400	0.475	0.427	0.620
	hit@5	0.330	0.432	0.325	0.373	0.612
	hit@1	0.284	0.275	0.475	0.267	0.452
KG²RAG	hit@3	0.284	0.352	0.475	0.480	0.496
	hit@5	0.341	0.530	0.500	0.533	0.484
	hit@1	**0.330**	**0.405**	0.600	**0.440**	**0.652**
MSGR	hit@3	**0.511**	**0.568**	**0.875**	**0.547**	**0.795**
	hit@5	**0.534**	**0.624**	**0.850**	**0.640**	**0.787**

在 LIRICAL 数据集上，MSGR 的性能提升相对较少，这一现象与数据集的固有特性密切相关。LIRICAL 数据集涵盖 252 种罕见疾病，但病例分布极度稀疏，平均每种疾病仅 1.47 个病例，导致实体关联密度显著低于其他数据集。这种稀疏性限制支链推理扩展的效果。低连通性导致多跳路径存在概率骤降，使推理链探索深度不足，易发生语义漂移，而置信度累积衰减进一步削弱长路径的可靠性。MSGR 双通道设计在当支链扩展因稀疏性失效时，精确子图匹配可以通过局部结构完整性提供降级保障，动态权重调整机制自动提升子图通道贡献度，形成稀疏环境与稠密环境都能够增效的智能切换策略，从而在极端数据分布下仍维持方法稳定性。MSGR 并非单一依赖推理深度，而是通过结构互补适应不同数据分布。

为了评估通用医疗咨询能力，从 PUMCH 数据集中选择数据，创建会诊病例评估数据集。PUMCH 数据集是从北京协和医院的真实数据中经过严格的数据清洗，去除不相关的信息，保证数据集的高质量和时效性[16]。这些病例涵盖多种罕见病场景，综合反映了罕见病诊断的的多样性和复杂性。基于这些案例，大语言模型生成对数据集问题的响应，并通过以下指标评估其医疗咨询能力：全面性、准确性、实用性和总体有效性。每个指标的最大得分为 10 分，并为每个指标计算模型在所有情况下的平均得分。全面性：涵盖与问题相关的所有关键方面，不遗漏重要信息。准确性：能根据给定的问题回答密切相关、准确的信息，帮助患者全面了解疾病信息，减少患者的误解和偏见。实用性：能根据患者病情分析解决方案，灵活应对不同情况，为临床诊疗提供切实可行的建议。整体有效性：有效整合患者信息，避免无关分析，为诊疗过程提供有价值的支持，提高诊断效率和有效性。

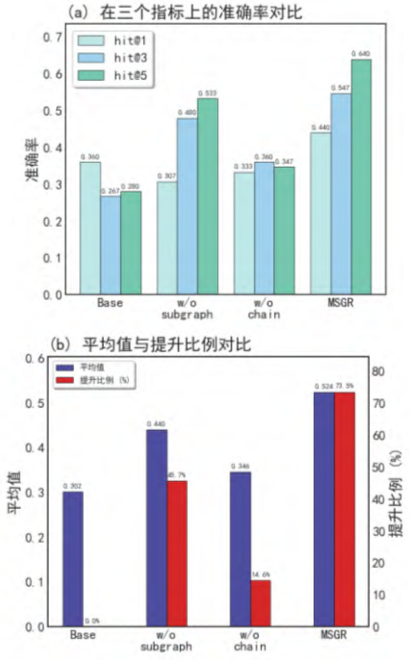
为确保客观性，使用 GPT-4o 作为评估工具，以自动评估会诊输出并根据这些标准对其进行评分，结果见表 2。本文的方法显著增强了模型的医疗会诊能力，在所有指标上均优于其他模型。这些结果表明，本文的多结构图检索方法在医疗会诊任务中具有优异的性能。根据结果评估，本文的方法大幅提高了大语言模型对于回答医疗问题的全面性和准确性，有效增强了大语言模型的医疗咨询能力，同时也凸显了其在现实医疗场景中的实际应用潜力。相比于其他基线方法，本文的方法在对于用户问题的回复上准确性更高，检索的信息更为全面且实用。

表 2　在医疗咨询能力上不同方法的比较

Tab. 2 Comparison of different methods in medical consultation ability

模型	全面性/分	准确性/分	实用性/分	总体有效性/分	平均分/分
LightRAG	7.74	8.99	5.97	7.49	7.55
Base	7.95	9.01	6.09	7.56	7.65
ToG	7.89	8.93	5.98	7.47	7.57
MindMap	8.01	8.64	5.92	7.56	7.53
MSGR	**8.66**	**9.02**	**7.44**	**8.37**	**8.37**

为了评估本文框架各个模块的有效性，进行消融实验，在 PUMCH 数据集上进行对比，结果如图 2 所示。



注：Base 为仅检索基于实体的三元组，w/o subgraph 为去掉精确子图匹配，w/o chain 为去掉支链推理扩展，MSGR 为完整多结构图检索架构。

图 2　消融实验

Fig. 2　Ablation study

由图 2（a）可知，本文方法每一部分设计都是有效的，相比于 w/o subgraph 方法、w/o chain 方法、Base 方法，本文方法平均准确率分别提升 10.2%、19.6%、24%。由图 2（b）可知，本文结合多种结构的方法平均准确率最高，并且相比于基于实体三元组的 Base 方法，单独精确子图匹配平均准确率提升 14.6%，单独支链推理扩展平均准确率提升 45.7%。本文的多结构图检索方法平均准确率提升 73.5%。

为系统阐释 MSGR 方法的优越性，本文将其与现有主流检索范式进行对比，结果见表 3。传统方法在此案例中均暴露其固有局限性：基于三元组的检索导致语义关联断裂，基于推理链的方法因聚焦单一路径而忽视关键证据，而基于子图的方法则引入大量无关信息干扰大语言模型决策。MSGR 通过其双通道检索与融合机制，精准地弥补了上述各项缺陷，为生成可靠诊断奠定了坚实的知识基础。

为评估不同检索方法的有效性，构建针对普拉德-威利综合征（Prader-Willi Syndrome, PWS）的诊断案例。该案例的用户查询含有一系列症状，包括强迫行为、言语语言发育迟缓、吮吸无力、产后生长迟缓及婴儿期肌张力低下。案例预设的正确诊断结果为 PWS。评估的重点在于考察各方法所检索的知识在辅助大语言模型生成准确且高置信度诊断方面的效用。本案例的核心挑战在于，所列举的症状在多基因遗传病中普遍存在，因此对检索知识的精准性与全面性提出了较高要求。该案例将用于下文对不同检索范式的对比分析。

表 3　不同检索方法在 PWS 案例上的对比分析

Tab. 3　Comparative analysis of different retrieval methods in PWS case

检索方法	检索内容	核心缺陷	MSGR 的解决方案与效果
基于三元组	(PWS，有症状，强迫行为)，(PWS，有症状，肌张力低下)，(Angelman 综合征，有症状，言语迟缓)...（离散、无关联的三元组）	图结构丢失：孤立的事实无法传递这些症状是 PWS 特征性集群的信息，增加了误诊（如天使综合征）的风险。	精确子图匹配构建了一个局部连通子图，保留了完整的疾病-症状簇结构。使大语言模型能识别出这是综合征，极大降低了错误。
基于推理链	吮吸无力→导致→喂养困难→与...相关→PWS（一条高置信度的线性推理路径）	检索焦点过窄：可能遗漏关键证据（如强迫行为），导致证据链不完整，缺乏诊断置信度。	支链推理扩展执行多跳探索，捕获到所有关键症状的隐式关联。其检索结果是一个证据网络，确保无一遗漏。
基于子图	一个以生长迟缓为中心的大规模子图，包含 PWS、特纳综合征、努南综合征等数十种疾病及上百个相关症状/基因。（庞大且嘈杂的子图）	信息冗余：大量无关信息（如其他疾病的症状）稀释了大语言模型的注意力，可能导致答案偏离主题或产生幻觉。	加权融合与关系去重，将输出精炼成一份简洁的结构化上下文。大语言模型的注意力完全集中在最相关信息（PWS）上，消除了噪声干扰。
MSGR	精炼、集成并去重的、高度特异于 PWS 的实体和关系集合（详见下表 5）。	**未发现**	双通道协同同时解决了所有三个局限性，提供了完整、精确、无冗余的知识，用于可靠诊断。

MSGR 方法的最终优势体现于其输出成果的质量。MSGR 并非简单返回原始数据，而是生成一份经

过深度加工、组织有序的结构化上下文（表 4），其输出清晰区分了实体与关系，并标注了知识来源通道。此举不仅极大提升了大语言模型的理解与推理效率，还显著增强了生成过程的可解释性，最终引导大语言模型输出如案例所示的高置信度准确诊断。

表 4　针对 PWS 案例中 MSGR 生成的结构化上下文

Tab. 4　Structured context generated for MSGR in PWS case

类型	编号	内容	来源
实体	E1	普拉德-威利综合征 (疾病)	子图通道
	E2	强迫行为 (症状)	子图通道
	E3	婴儿期肌张力低下 (症状)	子图通道
	E4	言语发育迟缓 (症状)	支链通道
	E5	吮吸无力 (症状)	支链通道
	E6	产后生长迟缓 (症状)	支链通道
关系	R1	(E1) 有症状 (E2)	子图通道
	R2	(E1) 有症状 (E3)	子图通道
	R3	(E1) 有症状 (E4)	支链通道
	R4	(E1) 有症状 (E5)	支链通道
	R5	(E1) 有症状 (E6)	支链通道

针对双通道协同作用进行了定量分析，通过网格搜索方式，搜索范围为[0.0-1.0]，步长为 0.2，结果见表 5。参数 ω_i 取值为 0.6 时，在 PUMCH 上的 hit@1 指标到最优。在实际检索中，子图通道权重占比 0.6，以确保结构完整性优先；支链通道权重占比 0.4，以确保检索结果包含推理深度补充。

表 5　参数 ω_i 的定量分析

Tab. 5　Quantitative analysis of parameter ω_i

ω_i 取值	描述	评分/分
0.0	仅使用支链推理扩展	0.347
0.2	支链主导（权重 80%支链）	0.400
0.4	平衡偏支链（权重 60%支链）	0.413
0.6	最优平衡（权重 60%子图）	0.440
0.8	平衡偏子图（权重 80%子图）	0.400
1.0	仅使用子图匹配	0.387

为了保证实际应用中的体验，针对本文方法进行实时性检索测试，在 PUMCH 数据集上测试检索时长全部在 20 s 以内，平均时长为 16 s，针对不同的大语言模型服务，检索时长可能会有所波动。针对医疗诊断场景，特别是罕见病这种需要经过长时间多专家会诊的任务，经过统计，迭代次数都在 4 次以内。这个检索延迟基本能够满足实际应用场景的需求，不同 API 供应商提供的大语言模型服务延迟也可能会有所差距。

4　结　论

本文提出多结构图检索增强方法(MSGR)，通过精确子图匹配与支链推理扩展的双通道协同机制，有效解决了知识图谱融合中的结构丢失、焦点过窄与信息冗余问题。本方法在医疗咨询任务中显著提升模型性能：在可解释性方面，通过结构化上下文生成与路径来源标注，实现了决策过程的透明化与可追溯；在抗幻觉能力方面，凭借双通道检索的互补校验机制，在医疗咨询基准测试中使实用性指标提升 22.2%，有效减少了事实性错误的产生；在知识覆盖度方面，通过支链推理扩展的多跳探索，全面性指标提升 8.9%，显著增强了隐性关联知识的挖掘能力。

本文方法通过关系去重算法与动态权重融合机制，建立了可解释、高覆盖的知识检索范式，为高风险领域提供了具备事实准确性与决策可追溯性的解决方案。消融实验结果表明，MSGR 相较于基准方法在准确率上提升显著，表明其在实际应用中的有效性。MSGR 为知识密集型场景的语言模型应用开辟了新途径，未来将向多模态知识融合方向扩展。

参考文献：

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[2] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[3] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[PP/OL]. V1. arXiv（2023-02-27）[2025-10-10]. https://arxiv.org/pdf/2302.13971v1.

[4] Anil R, Dai A M, Firat O, et al. PaLM 2 technical report[PP/OL]. V3. arXiv（2023-09-13）[2025-10-10]. https://doi.org/10.48550/arXiv.2305.10403.

[5] DeepSeek-AI. Deepseek-v3 technical report[PP/OL]. V2. arXiv（2025-02-18）[2025-10-10]. https://arxiv.org/pdf/2412.19437.

[6] Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling

language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[7] Ji Ziwei, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.

[8] Danilevsky M, Qian K, Aharonov R, et al. A survey of the state of explainable AI for natural language processing[PP/OL]. V1. arXiv（2020-10-01）[2025-10-10]. https://doi.org/10.48550/arXiv.2010.00711.

[9] Pan Shirui, Luo Linhao, Wang Yufei, et al. Unifying large language models and knowledge graphs: a roadmap[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3580-3599.

[10] 陈昱胤, 李贯峰, 秦晶, 等. 知识图谱的复杂逻辑查询方法研究综述[J]. 计算机科学, 2026, 53(2): 273-288.

[11] Hu Linmei, Liu Zeyi, Zhao Ziwang, et al. A survey of knowledge enhanced pre-trained language models[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(4): 1413-1430.

[12] Yasunaga M, Ren Hongyu, Bosselut A, et al. QA-GNN: reasoning with language models and knowledge graphs for question answering[PP/OL]. V5. arXiv（2022-12-13）[2025-10-10]. https://doi.org/10.48550/arXiv.2104.06378.

[13] Gao Yunfan，Xiong Yun，Gao Xinyu, et al. Retrieval-augmented generation for large language models: a survey[PP/OL]. V1. arXiv（2023-12-18）[2025-10-10]. https://doi.org/10.48550/arXiv.2312.10997.

[14] Zhu Chenguang，Xu Yichong，Ren Xiang, et al. Knowledge-augmented methods for natural language processing[C]//Proceedings of the sixteenth ACM international conference on web search and data mining. 2023: 1228-1231.

[15] 蒋献,王涵亦,杨诗婷,等. 基于大语言模型与知识图谱融合的多跳问答技术研究[J]. 电信科学，2025，41(11): 67-83.

[16] Ma Shengjie, Xu Chengjin, Jiang Xuhui, et al. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation[PP/OL]. V7. arXiv（2025-02-10）[2025-10-10]. https://arxiv.org/pdf/2407.10805

[17] Wen Yilin, Wang Zifeng, Sun Jimeng. MindMap: Knowledge graph prompting sparks graph of thoughts in large language models[PP/OL]. V5. arXiv（2022-12-13）[2025-10-10]. https://arxiv.org/html/2308.09729v5.

[18] Edge D, Trinh H, Cheng N, et al. From local to global: a graph rag approach to query-focused summarization[PP/OL]. V2. arXiv （2024-04-24）[2025-10-10]. https://arxiv.org/abs/2404.16130.

[19] Huang Haoyu, Huang Yongfeng, Yang Junjie, et al. Retrieval-augmented generation with hierarchical knowledge[C]//Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou: Association for Computational Linguistics, 2025: 6044-6060.

[20] Kang M, Baek J, Hwang S J. 2022. KALA: knowledge-augmented language model adaptation[C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics，2022: 5144–5167.

[21] Zhang Yifei, Wang Xintao, Liang Jiaqing，et al. Chain-of-knowledge: Integrating knowledge reasoning into large language models by learning from knowledge graphs[PP/OL]. V2. arXiv（2024-01-30）[2025-10-10]. https://doi.org/10.48550/arXiv.2407.00653.

[22] Jiang Xinke, Zhang Ruizhe, Xu Yongxin, et al. HyKGE: a hypothesis knowledge graph enhanced framework for accurate and reliable medical LLMs responses[PP/OL]. V2. arXiv （2023-12-26）[2025-10-10]. https://doi.org/10.48550/arXiv.2312.15883.

[23] Martínez-Demiguel C, Segura-Bedmar I, Chacón-Solano E, et al. The RareDis corpus: a corpus annotated with rare diseases, their signs and symptoms[J]. Journal of Biomedical Informatics, 2022, 125: 103961.

[24] Wang Guanchu, Ran Junhao, Tang Ruixiang, et al. Assessing and enhancing large language models in rare disease question-answering[PP/OL]. V7. arXiv（2024-08-15）[2025-10-10]. https://doi.org/10.48550/arXiv.2408.08422.

[25] Ronicke S, Hirsch M C, Türk E, et al. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study[J]. Orphanet Journal of Rare Diseases, 2019, 14(1): 69.

[26] Chen Xuanzhong，Mao Xiaohao, Guo Qihan，et al. RareBench: can LLMs serve as rare diseases

specialists?[C]//Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. 2024: 4850-4861.

[27] Töpel T, Scheible D, Trefz F, et al. RAMEDIS: a comprehensive information system for variations and corresponding phenotypes of rare metabolic diseases[J]. Human Mutation, 2010, 31(1): 1081-1088.

[28] Wei J, Wang Xuezhi, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[29] Guo Zirui, Xia Lianghao, Yu Yanhua, et al. LightRAG: Simple and fast retrieval-augmented generation[PP/OL]. V3. arXiv（2024-10-08）[2025-10-10]. https://doi.org/10.48550/arXiv.2410.05779.

[30] Zhu Xiangrong, Xie Yuexiang, LIU YI, et al. Knowledge graph-guided retrieval augmented generation[C]// Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Albuquerque: Association for Computational Linguistics, 2025: 8912-8924.