DOI:10.11784/tdxbz202507029

面向低空经济政策多跳问答的知识抽取策略

洪文兴[1,2]，李　宁[1]，李　俊[1]，黄燕燕[1]
（1. 厦门大学航空航天学院，厦门 361102；2. 福建省系统工程学会，厦门 361102）

摘　要：低空经济政策文本分析是典型的多跳问答任务，常面临推理链条复杂与长文本噪声干扰问题. 为此，提出一种子问题引导的细粒度知识抽取策略（SG-FKE），通过查询重写将复杂问题分解为语义明确的子问题，引导模型聚焦核心意图，并从长文本中抽取用于多跳推理的结构化三元组，去除冗余内容，构建清晰的推理路径，从而提升问答系统的抗干扰能力与推理准确性. 在公开多跳问答数据集 HotpotQA、2Wiki、MuSiQue 及自构建的低空经济政策问答数据集（LEMQA）上开展实验，实验结果显示，在 ChatGLM-9 B 模型上，SG-FKE 在 HotpotQA 的 F_1 和 E_M 分别达到 62.6%和 51.0%，在 2Wiki 上分别为 54.6%和 46.9%，在 MuSiQue 上分别为 48.8%和 29.9%，在 LEMQA 上 ROUGE 和 BERTScore-F_1 分别为 44.6%和 73.3%；在 Qwen-32 B 模型上各项指标进一步提升，均显著优于基线方法. 同时，实验还表明，SG-FKE 能够将原始上下文文本显著压缩至约 5%～15%的长度，有效减少冗余信息. 此外，案例分析显示，SG-FKE 不仅在逻辑构建、语义聚焦与知识整合方面具备优势，还能够在含有时效性条款的政策问答中生成与时间相关的子问题，抽取关键时间三元组，从而避免答案失真，提升了结果的可靠性. 消融实验进一步验证了子问题引导与三元组抽取机制在提升推理质量方面的关键作用. 整体结果表明，SG-FKE 能够增强大语言模型对低空经济政策长文本的多跳推理能力与时效性条款处理能力，为智能化政策问答提供了一种轻量化、可推广的技术方案，具有重要的理论意义与实践价值.

关键词：多跳问答；检索增强生成；知识抽取；低空经济政策
中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：0493-2137（2026）11-1197-11

Knowledge Extraction Strategy for Multi-Hop Question Answering over Low-Altitude Economy Policies

Hong Wenxing[1,2]，Li Ning[1]，Li Jun[1]，Huang Yanyan[1]
（1. School of Aerospace Engineering，Xiamen University，Xiamen 361102，China；
2. Systems Engineering Society of Fujian Province，Xiamen 361102，China）

Abstract：Analyzing low-altitude economy policy texts presents a typical multi-hop question answering（QA） challenge，often hindered by complex reasoning chains and noise from lengthy texts. To tackle these problems，a sub-question-guided fine-grained knowledge extraction strategy（SG-FKE）was developed. This strategy employed query rewriting to break down complex questions into clear，semantically focused subquestions，directing the model toward its core intent. Structured triples for multi-hop reasoning were extracted from extensive texts，redundant content was eliminated，and clear reasoning paths were constructed，thereby enhancing the robustness and accuracy of the QA system. Experiments were conducted on three public multi-hop QA datasets—HotpotQA，2Wiki，and MuSiQue—and a self-constructed dataset，the low-altitude economy multi-hop question answering dataset（LEMQA）. Results from the ChatGLM-9 B model show that SG-FKE achieves an F_1 score of 62.6% and an exact match rate of 51.0% on HotpotQA；54.6% and 46.9% on 2Wiki；48.8% and 29.9% on MuSiQue；and ROUGE and BERTScore-F_1 scores of 44.6% and 73.3% on LEMQA，respectively. All metrics further improve with the Qwen-32 B

收稿日期：2025-07-30；**修回日期**：2025-09-22.
作者简介：洪文兴（1980—　　），男，博士，教授.
通信作者：洪文兴，hwx@xmu.edu.cn.
基金项目：福建省创新战略研究计划资助项目（2024R0157）.
Supported by the Fujian Provincial Science and Technology Plan Innovation Strategy Research Project（No. 2024R0157）.

model, significantly outperforming baseline methods. In addition, our experiments show that SG-FKE can condense original context texts to approximately 5%—15% of their length, effectively minimizing redundancies. Case analyses demonstrate that SG-FKE excels in logical construction, semantic focus, and knowledge integration. It also generates time-related subquestions and extracts essential temporal triples in policy QA involving time-sensitive clauses, thus preventing answer distortion and increasing result reliability. Ablation studies confirm the importance of the subquestion guidance and triple extraction mechanisms in enhancing reasoning quality. Overall, our findings show that SG-FKE significantly improves the multi-hop reasoning and time-sensitive clause processing capabilities of large language models when working with lengthy low-altitude economy policy texts, thus offering a lightweight and generalizable technical solution for intelligent policy QA with significant theoretical and practical significance.

Keywords: multi-hop question answering; retrieval-augmented generation; knowledge extraction; low-altitude economy policy

低空经济是以低空飞行活动促进产业融合发展的综合性经济形态,广泛涵盖物流配送、精准农业、空中交通、城市管理及低空旅游等多个领域[1]. 近年来,在国家和地方政策的大力推动下,我国低空经济发展迅猛,基本形成了从中央到地方的多层级政策体系[2]. 低空经济政策文本内容复杂、篇幅较长,使其分析过程呈现出典型的多跳推理特征,面临推理链条复杂与长文本噪声干扰的挑战,亟需智能化的处理方案.

大语言模型(large language models, LLMs)依托海量预训练数据与深层神经网络的强大建模能力,在自然语言理解与生成等任务中展现出卓越性能,为政策文本分析开辟了新的技术路径[3]. 但其依赖静态预训练语料库,在面对动态更新的政策条款与专业术语时,存在知识边界瓶颈,难以应对训练数据范围之外的问题[4]. 检索增强生成(retrieval-augmented generation, RAG)[5]方法通过引入外部知识库扩展模型知识边界,但针对需多跳推理的复杂问题,RAG 对复杂查询的语义理解能力有限,难以构建精准的推理链条,导致回答质量不佳. 另一方面,在处理篇幅冗长、内容复杂的文档时,大语言模型容易受到文本噪声的干扰,即文中存在大量与目标问题无关的背景性描述、泛化性措辞或无效信息会削弱模型的信息聚焦能力,并诱发“幻觉”(hallucination)现象,从而影响回答的准确性与可信度. 为缓解上述问题,部分研究尝试将知识图谱引入检索增强生成流程,通过构建显式的推理路径提升模型的逻辑建模能力. 然而,在面对规模庞大的知识图谱时,此类方法往往面临检索和推理效率低、构建与更新成本高等问题[6].

针对低空经济政策多跳问答任务中的推理链条复杂和长文本噪声干扰问题,本文提出了一种子问题引导的细粒度知识抽取策略(subquestion-guided fine-grained knowledge extraction, SG-FKE). 具体而言,该策略首先将复杂多跳问题重写为语义明确的子问题,降低模型对原始问题的理解难度;随后,以这些子问题为提示,引导模型在原始政策文本中抽取与之对应的知识三元组,去除无关内容,形成一组轻量化、结构化的知识单元;最后,将所有抽取到的三元组与原始问题一并输入大语言模型,生成具有逻辑一致性与语义准确性的回答. 实验结果表明,SG-FKE 在多个公开多跳问答数据集(HotpotQA、2Wiki 和 MuSiQue)以及自构建的低空经济政策问答数据集(low-altitude economy multi-hop question answering dataset, LEMQA)上,在 F_1、E_M、ROUGE 和 BERTScore-F_1 等指标上均优于基线方法,为面向政策长文本的多跳问答任务提供了坚实的技术支撑.

本研究的主要贡献包括 3 点:首先,子问题引导机制增强了模型对复杂查询的理解能力,使知识抽取更聚焦于关键实体和关系,从而提升了三元组的抽取质量;其次,通过以细粒度三元组替代冗长文本输入,显著压缩了上下文长度,有效降低了生成阶段出现“幻觉”内容的风险;最后,抽取得到的三元组构建了显式的推理链条,为模型跨条款、跨段落的多步推理提供了结构化支持,显著提升了复杂问题回答的逻辑一致性与语义准确性.

1 多跳问答技术研究现状

1.1 大模型技术

在自然语言处理(natural language processing, NLP)领域,大模型近年来成为研究和应用的热点,其凭借卓越的性能和广泛的应用前景,推动着 NLP 技术快速发展. 大模型是指依托深层神经网络架构、包含数十亿至数千亿甚至万亿级参数,并在大规模语料上进行长周期自监督训练,从而获得跨任务泛化能力与高维语义建模能力的语言模型[7]. 具体来说,大模

型建立在 Transformer 架构之上，其中多头注意力层堆叠在深度神经网络中[8]. 大模型通常采用两阶段学习策略：预训练（pre-training）和微调（fine-tuning）. 预训练阶段，模型在大量无标注文本上进行学习，获取语言的通用表示；微调阶段，则针对特定任务的数据集调整模型参数，以适应具体的应用场景.

大模型在自然语言处理领域的应用广泛，涵盖文本分类、问答系统、机器翻译、文本摘要等诸多任务. 然而，在处理复杂任务如多跳问答任务时，大模型仍存在着局限性：一方面，大模型依赖内置知识，存在知识边界问题；另一方面，“幻觉”现象使得模型生成的信息可能不准确或不一致，进一步影响了答案质量[9]，特别是在使用思维链（chain-of-thought，CoT）等提示技术时，幻觉可能会逐渐累积[10]. 因此，将大模型与外部知识检索相结合成为研究的重要方向.

1.2　检索增强生成技术

为了解决大模型的知识边界问题，RAG 技术应运而生. RAG 将传统的信息检索技术与生成模型相结合，其核心原理如下：首先，知识库用于存储大量结构化或非结构化的知识数据，并对其中的文本进行切分，然后嵌入至向量数据库；接着，检索模块利用信息检索技术，根据用户查询从向量数据库中筛选出相关性最高的 Top-K 个文档或段落；最后，大模型基于用户输入及检索到的相关信息生成回答内容.

在 RAG 技术的发展历程中，研究者们为提升其性能提出了一系列优化策略. HyDE[11]使用查询转换的方式，利用查询生成伪文档辅助检索，从而使检索到的结果更加准确. RQ-RAG[12]则提供查询重写、分解和消歧功能来增强模型. BlendedRAG[13]使用语义搜索和基于混合查询的检索器来提高 RAG 准确性. Self-RAG[14]借助训练生成器，依据不同指令下的检索词符来判定是否检索，并评估检索文档的相关性与质量. 此外，部分学者尝试将 RAG 与微调相结合，Zhang 等[15]提出检索增强微调训练策略，训练模型忽略无关文档，直接引用相关文档内容作答，在多个数据集上取得显著成果. 另外，早期的 RAG 方法多依赖静态、有限的语料库进行检索，随着技术的持续进步，研究者开始探索运用大规模网络搜索扩展检索结果，这种方式不仅丰富了信息的多样性，还提升了检索的准确性和时效性[16].

尽管 RAG 的优化策略在提升大模型生成质量方面成效显著，但在处理多跳问题任务时，仍存在一定局限性：传统 RAG 模型在处理文本时，往往仅能捕捉基于语义相似性的信息，难以挖掘仅靠语义相似性无法全面体现的结构化关系知识；当 RAG 模型将多个文本片段串联作为生成提示时，常出现内容重复叙述的情况，导致生成文本上下文冗长复杂，降低了信息有效性；RAG 缺乏全局视角，仅依据检索到的上下文信息生成回答，在需要广泛上下文理解和综合分析的任务中，难以达到理想效果[17].

1.3　融合检索增强生成的知识图谱问答技术

为进一步突破上述局限，研究者尝试将知识图谱（knowledge graph）引入 RAG 框架，借助结构化知识优化检索与生成过程. 知识图谱作为一种结构化的知识表达方式，能够显式建模实体及其关系，为问答任务提供清晰的推理路径和语义关联. 与传统的检索方法不同，知识图谱通过节点和边构建图结构，使模型具备多跳推理和跨文档关系整合能力，成为多跳问答任务的重要工具.

GraphRAG[18]创新性地将知识图谱与 RAG 相结合，利用图结构实现实体关联检索和路径推理，使模型在生成过程中能够灵活调用知识图谱中的节点及关系信息，从而有效增强复杂推理任务的生成效果. 在实际应用中，GraphRAG 技术已得到广泛应用和扩展. Medical GraphRAG[19]专注于医疗领域，通过创新的三重图构建和 U-Retrieval 技术，提高了大模型在基于证据的医疗响应中的表现，并在多个医疗问答和健康核查基准测试中展现出卓越的准确性和可靠性. Yi 等[20]提出一种基于分布式文件系统的工业级知识图谱管理系统 KGFabric，通过语义增强的可编程图模型，实现跨领域知识融合与大规模关系的秒级分析，为大规模知识图谱的高效存储与计算提供了新范式. KAG[21]把知识图谱作为核心组件，通过 LLMs 友好的知识表示、逻辑形式引导的混合推理与知识对齐机制，显著提升了专业领域问答的准确性与逻辑严谨性，已在政务问答和医疗问答等场景中取得显著成效.

因此，目前多跳问答任务的主流研究方法存在以下亟待解决的问题：生成式大语言模型在缺乏外部知识支持时，易生成虚构或不准确信息，严重影响答案的可靠性；现有检索增强生成方法过度依赖语义相似性检索，容易引入冗余信息，且缺乏实体关系推理与跨段落信息整合能力；融合检索增强生成的知识图谱问答技术有助于解决多跳推理问题，但大规模知识图谱的构建与维护成本高昂. 针对这些挑战，本文提出子问题引导的细粒度知识抽取策略，在无需构建完整知识图谱的前提下，通过子问题驱动的结构化三元组抽取机制，有效压缩原始长文本并构建清晰的推理链条，从而显著提升大模型在多跳政策问答任务中的语义聚焦与推理一致性. 该方法在保持生成灵活性的

同时，兼顾了效率、准确性与可扩展性，为专业领域复杂问答任务提供了一种轻量、可行的新路径.

2 子问题引导的细粒度知识抽取策略

本方法的整体流程如图 1 所示. 首先，对用户提出的原始复杂问题进行查询重写，将其拆解为若干语义明确、结构清晰的子问题；随后，结合原始问题与子问题，引导大语言模型在检索到的上下文中抽取形式为[实体 1，关系，实体 2]的知识三元组；最后，将所有抽取得到的三元组与原始问题一并输入模型，生成语义准确、逻辑连贯的最终答案.

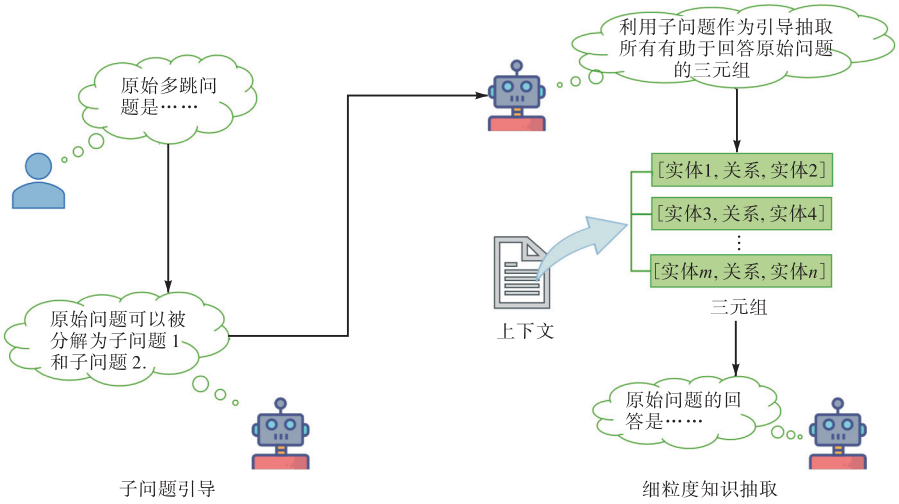


图 1 子问题引导的细粒度知识抽取策略流程

Fig.1 Workflow of SG-FKE

2.1 子问题引导策略

查询重写(query rewriting)是检索增强生成系统中常用的一种检索前预处理策略，旨在缓解用户查询中存在的表述不清、语义模糊等问题. 通过对原始问题进行语义层面的重写，使其更契合文档的语义空间，可有效提升检索阶段的信息相关性，并进一步改善大语言模型在生成阶段的回答质量. 目前已有多种典型的查询重写方法. 子问题策略将复杂问题拆解为若干更具体的子问题，以帮助模型更深入理解原始问题，从而提高检索的聚焦性与回答准确性. HyDE 通过生成假设文档，并利用其嵌入向量在向量空间中检索语义相近的真实文档，增强检索覆盖面. 回溯提示(backward prompting)通过将具体问题抽象化为更高层级的问题，引导模型把握背后的核心语义原则. 重写-检索-阅读(rewrite-retrieve-read)策略先使用大语言模型对原始问题进行改写，再执行信息检索与答案生成操作. 迭代检索与生成(iterative retrieval and generation)则通过迭代执行“检索增强生成”与“生成增强检索”，实现模型输出的逐步优化.

在本研究中，针对结构复杂、推理链条较长的多跳类问题，如图 2 所示，设计了一种基于少样本示例(few-shot)的查询重写策略，引导大语言模型将原始复杂问题拆解为两个语义清晰、结构明确的子问题.

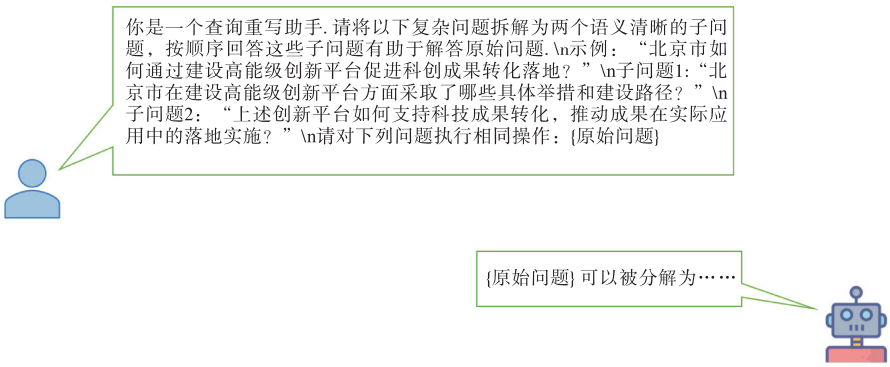


图 2 查询重写提示词

Fig.2 Prompt for query rewriting

需要强调的是，与传统的子查询策略不同，本文所生成的子问题并不直接用于独立检索或分步作答，而是在知识抽取阶段作为语义引导提示，帮助大语言模型在冗长文档中聚焦关键实体与关系，从而提升三元组抽取的相关性与准确性.

2.2　细粒度知识抽取策略

大语言模型具备一定的语义反思与信息筛选能力，能够在处理输入内容时识别并排除与问题无关的文本，从而降低冗余信息对生成结果的干扰，提升回答的准确性与稳定性. 同时，知识图谱通过构建显式的实体与关系结构，对复杂问题中的推理路径进行结构化表示，在多跳问答任务中展现出优异的推理与信息整合能力. 本文提出的细粒度知识抽取方法，结合了大模型在语义反思方面的优势，以及知识图谱在结构化推理中的能力，旨在从检索得到的政策文本中精准抽取能够回答问题的三元组. 该方法能够将冗长、结构松散的原始文本转化为信息密度高、结构紧凑的知识单元，在保留核心语义的前提下去除冗余内容，显著减轻模型的输入负担，并为后续的多跳推理奠定基础.

在具体实现过程中，将查询重写模块生成的子问题与原始问题一并输入大模型，引导其在指定上下文中抽取与原始问题紧密相关的知识三元组. 为此，设计了提示词如图 3 所示.

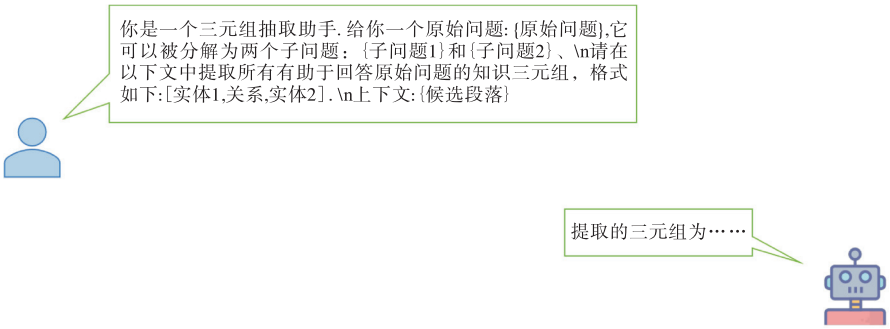


图 3　三元组抽取提示词

Fig.3　Prompt for triple extraction

此外，为提高抽取的一致性与格式规范性，为模型提供了 2、3 条高质量的三元组示例作为范例，并明确其输出应严格遵循[实体 1，关系，实体 2]的标准结构.

3　多跳问答任务实验验证与结果分析

3.1　实验设计

为全面评估本研究在处理复杂多跳问答任务中的实际效果，本文选取了 3 个具有代表性的公开数据集：HotpotQA[22]、2Wiki[23]和 MuSiQue[24]. 这些数据集均为多跳推理任务设计，问题通常需要经历 2 ~ 4 个逻辑推理步骤方可获得正确答案，且每个问题分别关联来自维基百科的 10、10 和 20 篇背景文档. 在实验过程中，本文将开发集按 7∶3 比例划分为验证集与测试集，最终测试集中分别包含 2 222 条、3 773 条和 725 条问题样本.

为了进一步验证所提方法在内容复杂、信息分布分散的政策文本分析场景中的适用性与泛化能力，本文自主构建了一个面向低空经济领域的多跳问答数据集，命名为 LEMQA. 该数据集聚焦于具有代表性的低空经济政策应用问题，旨在弥补现有通用问答数据集缺乏专业语境的不足.

为确保数据质量与覆盖广度，本文以国务院、民航局及各地方政府官网等权威平台为数据源，截至 2025 年 12 月 31 日共收集低空经济相关政策文件 283 份，涵盖“高质量发展实施方案”、“总体规划”、“专项措施”等多种类型，确保语料来源权威、内容结构多样、政策语义丰富. 基于上述政策语料，本文结合大语言模型与检索增强生成机制构建多跳问答样本，并引入人工审校以保证数据质量. 具体而言，对于每份政策文件，依据其结构层级与政策要素，设计 3 个涵盖跨段落、跨条款及跨实体的信息整合型问题. 在问答构造过程中，利用提示工程引导大语言模型生成候选问答对，并通过人工筛选和校对，评估其语义准确性、逻辑完整性及与政策文本的贴合度. 最终，保留 810 条语义清晰、推理链条完整的问答对，形成低空经济政策问答数据集. 该数据集具有如下特点与优势.

(1) 面向真实政策语境：所有问答样本均来源于公开发布的政策文件，具有高度现实性与领域代表性.

(2) 聚焦多跳推理任务：问题需跨越多个段落与实体，考验模型对长文本信息的综合理解与推理

能力.

(3) 长文本干扰信息丰富:政策文本篇幅较长,包含大量背景描述与冗余信息,提升了推理难度,有助于评估方法在复杂语境下的信息聚焦与抗噪声干扰能力.

为系统评估本研究在多跳问答任务中的性能表现,本文在 3 个公开多跳问答数据集上采用 F_1 分数与 E_M 作为主要评估指标,用于衡量模型预测答案与参考答案之间的匹配程度. F_1 分数基于 token-level 的精确率与召回率计算其调和平均值,能够综合反映预测答案与标准答案在词级层面的重叠情况;E_M 则用于判断归一化后的预测答案是否与标准答案完全一致,是一种更为严格的匹配标准. 在 LEMQA 数据集上,为更全面评估生成文本的语言质量与语义贴合度,本文进一步引入 ROUGE 与 BERTScore-F_1 作为评价指标. ROUGE 基于 n-gram 重叠或最长公共子序列计算生成文本与参考文本之间的词级匹配程度,能够反映文本重叠特征;BERTScore-F_1 则基于预训练语言模型生成的上下文嵌入表示,通过计算词向量之间的余弦相似度来衡量语义匹配程度,从而更有效地捕捉深层语义一致性.

本文选取 4 种具有代表性的基线方法进行对比评估:Naïve RAG 是 RAG 系统的标准基线,通过文本块嵌入与相似度检索实现高效直接的文段匹配;CoT 通过引导模型显式展示推理过程,提升复杂多跳问题的可解释性与回答准确率;RQ-RAG 借助查询重写与分解技术优化检索流程,增强模型对多跳问题的理解与检索能力;GraphRAG 在传统 RAG 的基础上引入结构化的图数据来增强大模型的上下文理解与推理能力.

3.2 实验结果

本文在 ChatGLM-9 B 与 Qwen-32 B 两款模型上验证了所提方法 SG-FKE 的有效性,实验结果如表 1 所示. 在 HotpotQA 数据集上,SG-FKE 在两种模型下均取得最佳表现,其中 ChatGLM-9 B 的 F_1 和 E_M 分别为 62.6% 与 51.0%,Qwen-32 B 的 F_1 与 E_M 分别为 65.3% 与 57.5%,显著优于基线方法. 在实体跳跃特征更为显著的 2Wiki 数据集中,SG-FKE 同样表现稳健,ChatGLM-9 B 的 F_1 和 E_M 为 54.6% 与 46.9%,Qwen-32 B 的 F_1 和 E_M 为 59.0% 与 53.0%,F_1 分数相较 GraphRAG 分别提升 2.6% 与 1.4%,进一步体现了方法在多跳推理场景下的优势. 在 MuSiQue 数据集中,SG-FKE 依然保持领先,Qwen-32 B 模型下 F_1 达 54.2%,E_M 达 35.0%,E_M 值较 RQ-RAG 和 GraphRAG 分别提升 6.0% 和 1.6%. 此外,在低空经济政策问答数据集 LEMQA 中,实验结果进一步验证了本研究在真实政策文本分析中的泛化能力. 在 ChatGLM-9 B 模型下,SG-FKE 在 ROUGE 和 BERTScore-F_1 上分别达到 44.6% 和 73.3%;在 Qwen-32 B 模型下,ROUGE 和 BERTScore-F_1 分别提升至 46.8% 和 76.1%,优于所有基线方法. 相比之下,CoT 策略在生成过程中通过分步推理引导模型,但在长文本场景下,推理链条可能出现中断或逻辑跳跃,从而增加“幻觉”出现的风险. RQ-RAG 方法的性能高度依赖子问题设计及其回答的准确性,因此可能受到查询构建偏差的影响. 尽管 GraphRAG 通过结构化知识增强了多跳推理能力,但在本实验设置下,其整体性能仍低于 SG-FKE,说明在无需构建完整知识图谱的前提下,细粒度结构化知识抽取同样能够有效支撑复杂多跳推理.

表 1 实验结果

Tab.1 Experimental results

模型	方法	HotpotQA		2Wiki		MuSiQue		LEMQA	
		F_1/%	E_M/%	F_1/%	E_M/%	F_1/%	E_M/%	ROUGE/%	BERTScore-F_1/%
ChatGLM-9 B	Naïve RAG	56.2	46.4	45.7	34.0	42.8	19.3	40.2	69.6
	CoT	57.2	47.0	46.8	36.0	44.3	22.1	42.2	71.4
	RQ-RAG	58.1	48.0	49.2	38.0	46.4	24.1	43.7	72.6
	GraphRAG	60.3	49.6	52.0	42.9	47.9	27.6	42.5	72.9
	SG-FKE	62.6	51.0	54.6	46.9	48.8	29.9	44.6	73.3
Qwen-32 B	Naïve RAG	58.9	48.4	49.4	37.1	46.9	23.0	42.8	71.6
	CoT	59.2	50.2	51.8	44.6	47.2	26.2	44.6	73.2
	RQ-RAG	60.4	52.7	54.3	48.0	48.9	29.0	45.1	74.7
	GraphRAG	63.6	55.1	57.6	51.2	53.1	33.4	46.3	75.7
	SG-FKE	65.3	57.5	59.0	53.0	54.2	35.0	46.8	76.1

为了量化 SG-FKE 压缩上下文的效果,本文统计了各数据集原始上下文与进行三元组抽取后的上下文平均字符数. 如图 4 所示,SG-FKE 能将原始文本显著压缩至约 5% ~ 15%的长度:在 HotpotQA 和 2Wiki 上,经过三元组抽取后的上下文平均字符数分别约为 629 和 463(ChatGLM-9 B)、668 和 505

(Qwen-32 B)，相比原始文本长度减少约 85%；在 MuSiQue 上，原始上下文超过 9 600 字符，而抽取后不足 500 字符，仅保留约 5%；在 LEMQA 上，抽取后的上下文平均字符数约为 569(ChatGLM-9 B) 和 614(Qwen-32 B)，同样大幅压缩原始文本. 进一步比较发现，Qwen-32 B 进行三元组抽取后的平均字符数略高于 ChatGLM-9 B，说明其在知识覆盖上更为细致，而 ChatGLM-9 B 的压缩更为紧凑. 整体来看，SG-FKE 在不同复杂度的数据集上均能显著去除冗余信息、保留关键知识单元，有效缓解大模型在长文本推理中面临的输入冗余与噪声干扰问题.

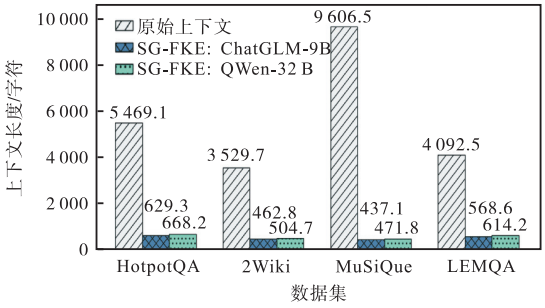


图 4　原始上下文与 SG-FKE 上下文长度对比

Fig.4　Comparison of original context and SG-FKE extracted context lengths

为了更直观地展示本研究在低空经济政策问答场景中的应用效果与优势，选取 LEMQA 中的一个案例进行深入分析. 如图 5 所示，问题“为支持低空物流发展，深圳市针对小型无人驾驶航空器运营主体设定了哪些奖励机制？”检索到《深圳市交通运输局 < 深圳市支持低空经济高质量发展的若干措施 > 实施细则》等相关政策文件. 针对该复杂问题，首先将其分解为两个子问题：①“深圳市针对开展低空物流的小型无人驾驶航空器运营主体，设置了哪些类型的奖励？”；②“深圳市针对开展低空物流的小型无人驾驶航空器运营主体，各类奖励的具体申请条件和奖励标准是什么？”. 在子问题的语义引导下，模型从政策文本中抽取出与问题直接相关的结构化三元组，这些三元组作为高密度的知识单元输入大语言模型，用以替代冗长的原始政策文本. 模型基于这些精炼信息生成语义准确的最终回答. SG-FKE 能够以子问题作为引导提示，从原始长文本中提取结构化三元组，构建清晰的推理路径，有效增强大模型在低空经济政策问答任务中的准确性和稳定性. 该方法在逻辑结构构建、语义聚焦和知识整合方面均表现出明显优势.

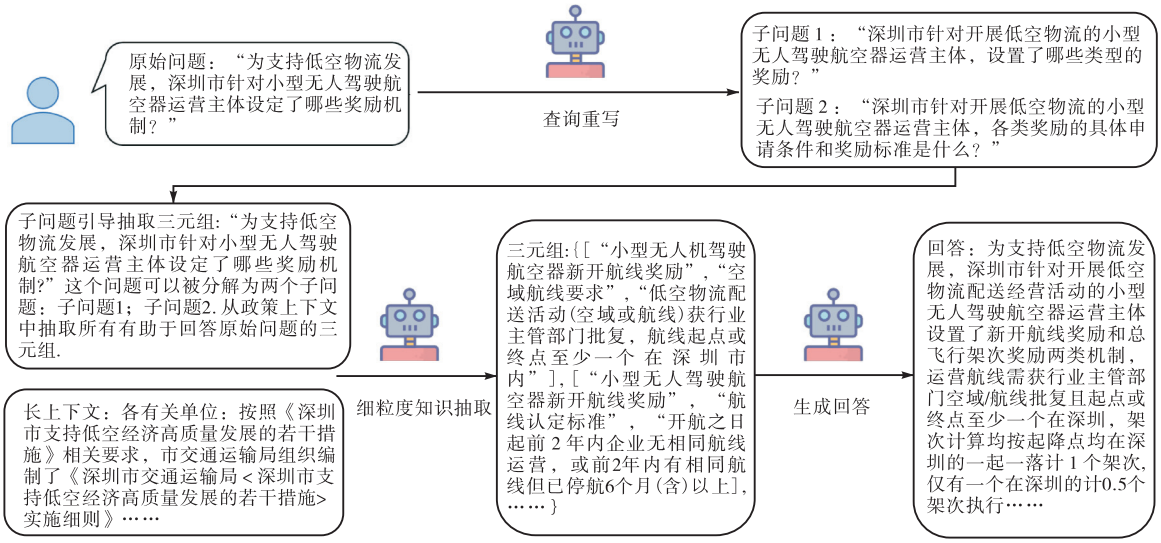


图 5　低空经济政策问答示例

Fig.5　Example of a low-altitude economy policy question answering

在低空经济政策问答中，部分文件具有明确的时效性约束，若未能正确识别其时间边界，可能导致答案不可靠或存在误导. 子问题引导的细粒度知识抽取策略能够有效缓解这一问题，其核心在于在查询重写阶段识别问题中含有的时间约束，并生成相应的子问题，以引导模型关注政策条款的时效性. 如图 6 所示，对于问题“雄安新区针对低空经济核心产品企业的扶持政策在有效期内需要满足哪些条件？”，查询重写将其分解为两个子问题：①“雄安新区针对低空经济核心产品企业的扶持政策的有效期是什么时候？”；②“在该有效期内，企业需要满足哪些条件才能获得扶持？”. 通过上述子问题的引导，模型能够在后续知识抽取过程中优先识别与有效期相关的实体和关系，抽取关键三元组，从而形成同时满足时效约束与条件要求的完整答案. 该过程表明，SG-FKE 在处理含有时效性约束的政策问答任务时，能够显式

构建时间相关的子问题,从而提升问答结果的准确性与可靠性.

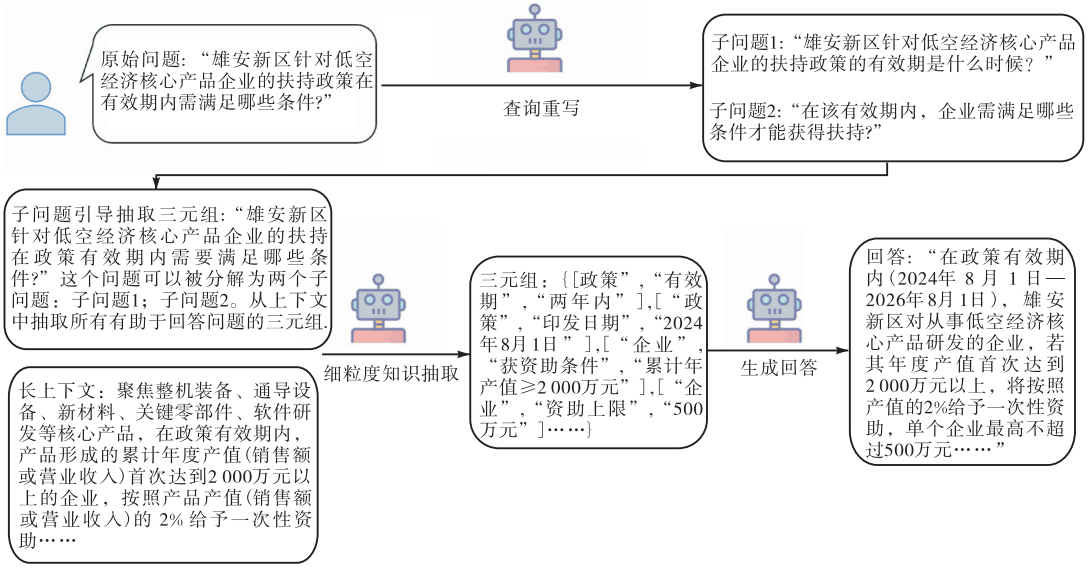


图 6 低空经济政策时效性条款问答示例

Fig.6 Example of question answering on time-sensitive clauses in low-altitude economy policy

3.3 关键模块消融实验

为进一步验证子问题引导机制和细粒度知识抽取模块在本方法中的核心作用,本文分别去除上述两个关键模块进行对比实验. 结果如表 2 和表 3 所示.

表 2 去除子问题引导实验结果

Tab.2 Experimental results without subquestion guidance

数据集	指标	ChatGLM-9 B		Qwen-32 B	
		无子问题引导	SG-FKE	无子问题引导	SG-FKE
HotpotQA	F_1/%	59.9	62.6	63.4	65.3
	E_M/%	49.7	51.0	54.5	57.5
2Wiki	F_1/%	51.7	54.6	56.7	59.0
	E_M/%	44.1	46.9	50.3	53.0
MuSiQue	F_1/%	46.1	48.8	51.5	54.2
	E_M/%	26.3	29.9	31.9	35.0
LEMQA	ROUGE/%	42.2	44.6	45.1	46.8
	BERTScore-F_1/%	72.3	73.3	74.9	76.1

表 3 去除三元组抽取实验结果

Tab.3 Experimental results without triple extraction

数据集	指标	ChatGLM-9 B		Qwen-32 B	
		无三元组抽取	SG-FKE	无三元组抽取	SG-FKE
HotpotQA	F_1/%	57.1	62.6	60.3	65.3
	E_M/%	47.5	51.0	51.0	57.5
2Wiki	F_1/%	48.8	54.6	52.4	59.0
	E_M/%	39.1	46.9	46.2	53.0
MuSiQue	F_1/%	43.1	48.8	46.3	54.2
	E_M/%	23.0	29.9	28.1	35.0
LEMQA	ROUGE/%	39.9	44.6	42.4	46.8
	BERTScore-F_1/%	69.7	73.3	72.7	76.1

去除子问题引导模块后,模型在多个数据集上的表现均出现不同程度下降,表明子问题引导在多跳推理过程中发挥了重要作用. 在 HotpotQA 数据集上,ChatGLM-9 B 的 F_1 由 62.6%降至 59.9%,E_M 由 51.0%降至 49.7%;Qwen-32 B 的 F_1 从 65.3%降至 63.4%,下降 1.9%,E_M 从 57.5%降至 54.5%,下降 3.0%. 在 2Wiki 数据集上,ChatGLM-9 B 的 F_1 从 54.6%降至 51.7%,E_M 从 46.9%降至 44.1%,分别下降 2.9%和

2.8%；Qwen-32 B 的 F_1 从 59.0%降至 56.7%，E_M 从 53.0%降至 50.3%，分别下降 2.3%和 2.7%. 在 MuSiQue 数据集上，ChatGLM-9 B 的 F_1 由 48.8%降至 46.1%，E_M 由 29.9%降至 26.3%；Qwen-32 B 的 F_1 从 54.2%降至 51.5%，E_M 从 35.0%降至 31.9%. 在低空经济政策问答数据集上，去除子问题引导后生成质量出现下降. ChatGLM-9 B 的 ROUGE 由 44.6%降至 42.2%，BERTScore-F_1 由 73.3%降至 72.3%；Qwen-32 B 的 ROUGE 由 46.8%降至 45.1%，BERTScore-F_1 由 76.1%降至 74.9%.

上述结果表明，去除子问题引导模块削弱了模型在多跳问答任务中理解和组织信息的能力，影响了最终答案的准确性，尤其在推理链条较长、跳数较多的语境下表现更为突出. 这充分证明了子问题引导机制在帮助模型理解复杂问题及聚焦关键知识三元组中的重要性.

去除细粒度知识抽取模块后，整体性能下降更加显著. 在 HotpotQA 数据集上，ChatGLM-9 B 的 F_1 从 62.6%降至 57.1%，E_M 从 51.0%降至 47.5%；Qwen-32 B 的 F_1 和 E_M 分别从 65.3%降至 60.3%、从 57.5%降至 51.0%，下降幅度分别为 5.0%和 6.5%. 在 2Wiki 数据集上，ChatGLM-9 B 的 F_1 和 E_M 从 54.6%和 46.9%分别降至 48.8%和 39.1%；Qwen-32 B 从 59.0%和 53.0%分别降至 52.4%和 46.2%，呈现出一致的性能退化趋势. 在 MuSiQue 数据集上，ChatGLM-9 B 的 F_1 从 48.8%降至 43.1%，E_M 从 29.9%降至 23.0%，分别下降 5.7%和 6.9%；Qwen-32 B 的 F_1 和 E_M 则从 54.2%和 35.0%分别降至 46.3%和 28.1%，性能下降更为明显. 在 LEMQA 上亦观察到相同趋势. ChatGLM-9 B 在移除细粒度知识抽取后，ROUGE 从 44.6%降至 39.9%，BERTScore-F_1 从 73.3%降至 69.7%；Qwen-32 B 的 ROUGE 从 46.8%降至 42.4%，BERTScore-F_1 从 76.1%降至 72.7%.

由此可见，细粒度知识抽取通过去除冗余信息、提取有效知识，构建推理路径，显著增强了模型的抗噪声干扰与多跳推理能力. 移除该模块将严重削弱模型对内容复杂或上下文较长问题的推理表现.

综合两组消融实验结果可得，细粒度知识抽取是本研究的核心模块，通过构建知识三元组压缩冗余文本、保留关键事实，有效提升了模型的多跳推理和抗噪声干扰能力；子问题引导模块则增强了模型对复杂查询的语义理解能力，通过将原始问题拆解为子问题，明确了抽取目标并提升了语义聚焦性. 两者协同作用形成互补机制. 实验结果充分验证了本方法在复杂问答场景中的有效性，为提升大语言模型的推理能力与鲁棒性提供了可行路径.

4　结　语

本文针对低空经济政策多跳问答中的推理链条复杂与长文本噪声干扰问题，提出一种子问题引导的细粒度知识抽取策略，从查询理解和知识组织两个层面提升模型的推理和抗噪声干扰能力. 具体而言，该方法通过生成子问题，引导模型关注核心意图，抽取关键三元组，构建结构化知识推理路径，从而支撑复杂多跳推理任务. 实验结果表明，SG-FKE 在多个公开多跳问答数据集上的 F_1 和 E_M 指标均优于基线模型，在低空经济任务中也实现了 BERTScore-F_1 与 ROUGE 指标的显著提升. 同时，三元组抽取有效压缩了上下文长度，减少了输入冗余与噪声干扰. 进一步的案例分析显示，本研究不仅能够在低空经济政策问答场景中展现优势，还能够在含有时效性条款的政策问答任务中捕捉时间边界信息，生成更准确和可靠的结果. 总体而言，SG-FKE 在政策类多跳问答场景中具有良好的实用性与推广潜力.

参考文献：

[1] 刘先江，宋　丹，徐　政. 以低空经济打造新质生产力发展新引擎[J]. 北京航空航天大学学报(社会科学版)，2024，37(5)：134-144.
Liu Xianjiang，Song Dan，Xu Zheng. Creating a new engine for the development of new quality productive forces through low-altitude economy[J]. Journal of Beijing University of Aeronautics and Astronautics (Social Sciences Edition)，2024，37(5)：134-144(in Chinese).

[2] 孔得建，袁　泽. 低空经济政策法律体系的现状、经验与展望[J]. 北京航空航天大学学报(社会科学版)，2024，37(5)：85-95.
Kong Dejian，Yuan Ze. Current situation，experience and prospect of the policies and legal system of low-altitude economy[J]. Journal of Beijing University of Aeronautics and Astronautics(Social Sciences Edition)，2024，37(5)：85-95(in Chinese).

[3] Kramer M A，Leavens A，Scarlat A. Harnessing AI for efficient analysis of complex policy documents：A case study of Executive Order 14110[EB/OL]. https://arxiv.

org/abs/2406.06657，2024-06-10.

[4] Ren R，Wang Y，Qu Y，et al. Investigating the factual knowledge boundary of large language models with retrieval augmentation[C]//Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi，UAE，2025：3697-3715.

[5] Lewis P，Perez E，Piktus A，et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems，2020，33：9459-9474.

[6] 冯拓宇，李伟平，郭庆浪，等. 大语言模型增强的知识图谱问答研究进展综述[J]. 计算机科学与探索，2024，18(11)：2887-2900.
Feng Tuoyu，Li Weiping，Guo Qinglang，et al. Overview of knowledge graph question answering enhanced by large language models[J]. Journal of Frontiers of Computer Science and Technology，2024，18(11)：2887-2900(in Chinese).

[7] Kumar P. Large language models(LLMs)：Survey，technical frameworks，and future challenges[J]. Artificial Intelligence Review，2024，57(10)：260.

[8] 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. Long Beach，USA，2017：6000-6010.

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

[10] Wei J，Wang X Z，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.

[11] Gao L，Ma X G，Lin J，et al. Precise zero-shot dense retrieval without relevance labels[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics(Volume 1：Long Papers). Toronto，Canada，2023：1762-1777.

[12] Chan C M，Xu C P，Yuan R B，et al. RQ-RAG：Learning to refine queries for retrieval augmented generation[EB/OL]. https://arxiv.org/abs/2404.00610，2024-03-31.

[13] Sawarkar K，Mangal A，Solanki S R. Blended RAG：Improving RAG（retriever-augmented generation）accuracy with semantic search and hybrid query-based retrievers[EB/OL]. https://arxiv.org/abs/2404.07220，2024-08-04.

[14] Asai A，Wu Z Q，Wang Y Z，et al. Self-RAG：Learning to retrieve，generate，and critique through self-reflection[EB/OL].https://arxiv.org/abs/2310.11511，2023-10-17.

[15] Zhang T J，Patil S G，Jain N，et al. Raft：Adapting language model to domain specific RAG[EB/OL]. https:// arxiv.org/abs/2403.10131，2024-03-15.

[16] Fan W Q，Ding Y J，Ning L B，et al. A survey on RAG meeting LLMs：Towards retrieval-augmented large language models[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York，USA，2024：6491-6501.

[17] Liu N F，Lin K，Hewitt J，et al. Lost in the middle：How language models use long contexts[J]. Transactions of the Association for Computational Linguistics，2024，12：157-173.

[18] Edge D，Trinh H，Cheng N，et al. From local to global：A graph RAG approach to query-focused summarization[EB/OL]. https://arxiv.org/abs/2404.16130，2024-04-24.

[19] Wu J D，Zhu J Y，Qi Y L，et al. Medical graph RAG：Evidence-based medical large language model via graph retrieval-augmented generation[C]// Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics(Volume 1：Long Papers). Vienna，Austria，2025：28443-28467.

[20] Yi P，Liang L，Zhang D，et al. KGFabric：A scalable knowledge graph warehouse for enterprise data interconnection[J]. Proceedings of the VLDB Endowment，2024，17(12)：3841-3854.

[21] Liang L，Bo Z P，Gui Z K，et al. KAG：Boosting LLMs in professional domains via knowledge augmented generation[C]//Companion Proceedings of the ACM on Web Conference 2025. Sydney，Australia，2025：334-343.

[22] Yang Z L，Qi P，Zhang S Z，et al. HotpotQA：A data-

set for diverse，explainable multi-hop question answering[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels，Belgium，2018：2369-2380.

[23] Ho X，Nguyen A K D，Sugawara S，et al. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps[C]// Proceedings of the 28th International Conference on Computational Linguistics. Barcelona，Spain，2020：6609-6625.

[24] Trivedi H，Balasubramanian N，Khot T，et al. MuSiQue：Multihop questions via single-hop question composition[J]. Transactions of the Association for Computational Linguistics，2022，10：539-554.

（责任编辑：王晓燕）