可靠的推理路径：提取知识图谱用于 LLM 推理的有效指导

Yilin Xiao, Chuang Zhou, Qinggang Zhang, Bo Li, Qing Li, *Fellow, IEEE*, Xiao Huang

***Abstract*—大型语言模型（LLMs）通常由于缺乏背景知识和幻觉倾向而在知识密集型任务上表现挣扎。为了解决这些限制，结合知识图谱（KGs）与 LLMs 已被深入研究。现有的 KG 增强型 LLMs 专注于补充事实性知识，但在解决复杂问题上仍然困难。我们认为，优化事实之间的关系并将它们组织成逻辑一致的推理路径与事实知识本身同等重要。尽管 KGs 具有潜力，但从中提取可靠的推理路径面临以下挑战：图结构的复杂性及存在多个生成的路径，使得区分有用和冗余路径变得困难。为了解决这些挑战，我们提出了 RRP 框架来挖掘知识图谱，该框架结合了 LLMs 的语义优势与通过关系嵌入和双向分布学习获得的结构信息。此外，我们引入了一种重新思考模块，根据其重要性评估和优化推理路径。在两个公共数据集上的实验结果表明，与现有基线方法相比，RRP 实现了最先进的性能。此外，RRP 可以轻松集成到各种 LLMs 中，以增强它们的推理能力，以插件即用的方式。通过生成针对特定问题的高质量推理路径，RRP 为 LLM 推理提炼了有效的指导。**

***Index Terms*—Knowledge graph, reasoning path, knowledge-intensive task, KG-enhanced LLMs.**

I. 引言

LLM 阿尔格 语言模型（LLMs）被广泛应用于实际应用中，并在各种自然语言理解和生成任务中展示了卓越的能力 [1]–[4] ，这得益于它们从大规模语料库预训练中获得的强大语义能力。然而，LLMs 在知识密集型任务中仍然表现出幻觉倾向，特别是当它们遇到需要外部背景知识或涉及复杂多跳推理的问题时 [5], [6] 。这些幻觉源于它们的固有特性，即主要依赖于从大规模语料库中学习的统计关联，而不是对潜在逻辑关系的结构化理解。因此，它们可能会产生流畅但事实不准确或逻辑不一致的回答。最近的努力集中在两个主要范式上：使用领域特定数据对 LLMs 进行微调，以及采用检索增强生成（RAG）技术，这种技术在推理过程中动态地结合外部知识。然而，微调可能通过对数据中的虚假关联进行过拟合而引入新的幻觉，而不是以一种可推广的方式提高事实一致性 [7]–[10] 。此外，微调计算成本高，无法应用于实时问题。检索增强生成 [11]–[14] 通过在推理时为大型语言模型（LLM）补充外部知识，提供了一种有前景的替代方案。虽然这种方法在解决相对简单的查询（例如提供定义或澄清术语）方面有效，但通常无法处理更复杂的推理任务 [15] 。所检索的文档通常是非结构化和断开的，这使得模型难以构建一个连贯的、多步骤的推理链。为了缓解这些问题，最近的研究探索了将 LLM 与外部知识源（特别是知识图谱（KG））整合的方案，知识图谱提供了实体及其关系的结构化和事实性表示 [16]–[18] 。经过 KG 增强的 LLM 通过基于实体连接补充相关背景知识，显示出在事实召回和问答中的性能提升。

Yilin Xiao, Chuang Zhou, Qinggang Zhang, Bo Li, Qing Li and Xiao Huang are with The Hong Kong Polytechnic University, Hong Kong SAR, China (e-mail: { yilin.xiao, chuang-qqzj.zhou, qinggangg.zhang } @connect.polyu.hk, comp-bo.li@polyu.edu.hk, csqli@comp.polyu.edu.hk, xiao.huang@polyu.edu.hk.).

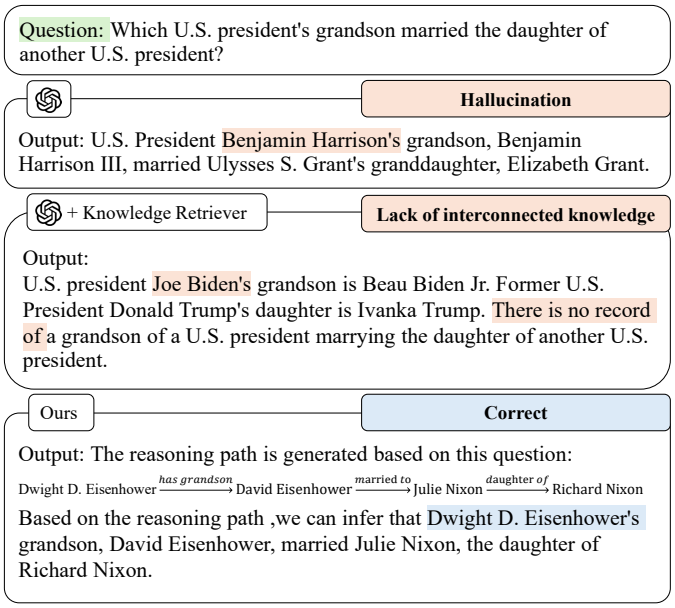


Fig. 1. 大型语言模型在回答复杂问题时经常会出现幻觉。将大型语言模型与知识检索器结合可以获取事实，但在处理需要对相互关联知识进行推理的问题时会遇到困难。我们的方法通过推理路径连接知识，能够为复杂问题提供准确的解决方案。

然而，虽然我们承认这样的图可以为知识注入带来显著的好处，但我们发现现有的很多工作 [19] 主要集中在检索适用的内容，而忽略了知识之间的逻辑。除了简单的事实检索外，合成逻辑一致且与上下文相关的推理路径的能力对于实现强大的 LLM 推理至关重要。例如，回答一个复杂的问题可能需要通过推理步骤结合多个证据，这些步骤超越了如图 1 中所示的表面级的实体连接。知识图谱不仅仅是提供知识单元，它们还包含了推理的逻辑。仅仅检索这些独立的知识单元不足以解决复杂的现实问题，因为这类挑战需要超越孤立事实的整合处理。在复杂场景中有效解决问题依赖于跨单元关系的系统组织和可靠推理路径的构建，这二者共同促进了一致的知识遍历和逻辑推理。

因此，我们被激励去组织知识单元之间的关系，并促进可靠推理路径的生成。然而，由于两个主要原因，这并非易事。(1) 图的复杂性：这些图的错综复杂的连接和特征常常使得关键信息难以识别。尽管 LLMs 具有强大的语义能力，但它们无法充分利用复杂结构信息。例如，需要多跳推理路径才能获得准确答案的问题对 LLMs 来说可能具

有挑战性。(2) 推理路径的准确性：我们观察到现有方法生成了大量的推理路径直接输入到 LLMs 中。这种方法是不合理的，因为正确的推理路径可能在众多路径中变得模糊，从而降低其有效性。此外，矛盾的推理路径可能会在 LLMs 中引入混淆。处理大量路径也降低了效率。

为了解决这些挑战并提炼出有效的指导原则用于 LLMs，我们提出了可靠推理路径（RRP），这一新颖的框架将 LLMs 和知识图谱（KGs）结合起来，以生成全面且高质量的推理。RRP 利用关系嵌入和双向分布学习来提取结构上连贯且与上下文相关的推理路径。为了进一步提高推理的可靠性，我们引入了重新思考模块，该模块根据生成路径对解决问题的贡献进行评估和排行。然后，得到的高质量推理路径被用来提炼为 LLMs 提供精确且可解释的指导，大大提升了它们在推理任务中的表现。

我们的主要贡献如下：

- 我们提出了一种新颖的框架，可靠推理路径（RRP），它紧密地整合了大型语言模型（LLMs）和知识图谱（KGs），以促进结构化推理。RRP 不仅关注于检索事实性知识，还关注于将其组织成针对特定问题的连贯推理路径。
- 我们开发了一个强大的推理路径生成模块，该模块结合了大型语言模型（LLMs）的语义能力和知识图谱（KGs）的结构先验。该模块结合了关系嵌入以捕捉实体之间的潜在连接，并采用双向分布学习以确保生成路径的一致性和完整性，从而实现更准确的多跳推理。
- 我们引入了一个重新思考模块，该模块从结构和语义的角度对生成的推理路径进行关键分析，在此基础上根据相关性与逻辑连贯性选择最具信息性的路径用于 LLM 推理，同时消除冗余和无效的路径。
- 在两个基准数据集上的实验表明，我们的框架在推理任务中表现出色，使用仅 7B 参数的 LLM 时，优于所有最先进的方法。我们还评估了模型中每个模块的效果以及与各种 LLM 的即插即用特性。

II. 相关工作

为了定制适合知识密集型任务的大型语言模型（LLMs），研究人员提出了多种方法以增强 LLMs 的推理能力，主要分为两大类：不使用知识图谱（KG）增强的 LLMs 和知识图谱增强的 LLMs。

A. 增强 LLMs 无需知识图谱

早期的尝试集中在微调技术以增强大型语言模型的推理能力 [20] 。然而，通过微调整合新知识会导致模型生成新的幻觉，甚至出现灾难性遗忘 [7] ，尤其是在新知识与之前学到的内容相矛盾时 [21] 。

最近，检索增强生成模型已经被广泛研究，以通过文本语料或在线来源的外部知识来增强 LLM。[22] 然而，这些方法在知识密集型任务中面临挑战：(*i*) 这些 RAG 文档可能具有不同质量、准确性和完整性，导致检索的知识中可能存在不一致或错误。[23] (*ii*) 缺乏显式关系和结构化组织限制了 RAG 模型的推理能力，因为它们无法利用结构化的连接来推导新的见解或生成更符合上下文的响应。[15]

B. 知识图谱增强的大型语言模型

早期的研究采用了启发式的方法，在预训练或微调期间从知识图中向大语言模型注入知识。ERNIE [24] 在预训练阶段结合了实体嵌入，并将其与词嵌入对齐，鼓励模型更好地理解和推理实体。UniKGQA [19] 是首次尝试利用大语言模型的能力，在一个统一的框架中联合进行知识检索和多跳推理。

另一类工作关注于在推理时从知识图谱中检索相关知识以增强语言模型的上下文。通常，K-BERT [25] 使用注意力机制从知识图谱中选择与输入上下文相关的三元组，然后将这些三元组附加到输入序列中。最近，知识图谱提示已被深入研究用于将事实性知识整合到大型语言模型中。KD-CoT [26] 和 KG-CoT [18] 基于思维链的概念，引导大型语言模型通过逐步的推理过程，同时允许及时纠正错误推理。它们的事实性和忠实性通过外部知识图谱进行验证。RoG [17] 提出了一种规划-检索-推理框架，该框架结合了大型语言模型和知识图谱，用于更透明和可解释的推理。SubgraphRAG [27] 将轻量级的多层感知机与并行的三元组评分机制整合在一起，实现高效而灵活的子图检索。尽管这些方法是有效的，现有模型主要关注知识检索或生成的方法设计。它们直接将检索到或生成的知识输入到大型语言模型中进行推理，但这并不一定能为大型语言模型提供有效的指导。这是因为大型语言模型往往难以区分有效和冗余的知识。

III. 初步

- 知识图谱（KG）。KG 以图的形式表示结构化的知识，其中信息被编码为一组三元组：$\mathcal{G} = \left\{\left\langle e, r, e^{'} | e, e^{'} \in \mathcal{E}, r \in \mathcal{R} \right\rangle\right\}$，，其中 $\mathcal{E}$ 和 $\mathcal{R}$ 表示实体集和关系集，e 和 $e^{'}$ 是实体，而 r 对应于实体之间的一个有向关系。
- 推理路径。在知识图谱（KG）中的推理路径可以被视为关系路径 $\{r_1, r_2, \ldots, r_n\}$ 的实例，其中一系列实体和关系形成了一条路径。具体而言，一个推理路径可以表示为：$\gamma = e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} \ldots \xrightarrow{r_n} e_n$ 。
- 利用知识图谱进行 LLM 推理是一项基本的推理任务，其重点是通过利用存储在知识图谱中的结构化知识来回答自然语言问题。给定一个自然语言问题 q 和一个知识图谱 $\mathcal{G}$ ，这项任务的目标是首先准确识别出问题中提及或隐含的实体，然后通过推理知识图谱内的关系和实体来生成正确答案。

在这一部分中，我们提出了 RRP，它由三个关键组件组成：1）语义推理路径生成：该组件利用大型语言模型（LLMs）的强大语义能力来生成与知识图谱中给定问题语义相关的推理路径。2）结构推理路径生成：为了解决 LLMs 在捕捉图结构信息方面的局限性，特别是在多跳推理中，图中的结构关系可能与问题的语义不完全一致。我们利用关系嵌入和双向分布学习来挖掘知识图的结构信息。该组件生成的推理路径在结构上与问题一致，从而补充了 LLMs 的语义生成。3）重新思考模块：该模块通过消除冗余、重新排列路径并优先考虑最相关的路径来优化生成的推理路径。这一步骤显著改善了对 LLM 推理的有效引导。知识图谱中的实体是动态更新的，这在推理过程中对某个具体实体过于依赖时引入了不确定性。相比之下，推理路径代表了知识图谱中更稳定和结构化的关系。作为连接实体的关系序列，推理路径为基于知识图谱的推理任务提供了坚实的基础。因此，我们利用大型语言模型的强大推理能力来生成推理路径。该模块的目标是生成尽可能可靠的知识图谱推理路径。为此，我们最小化可靠推理路径的后验分布与先验分布之间的 KL 散度。给定一个问题

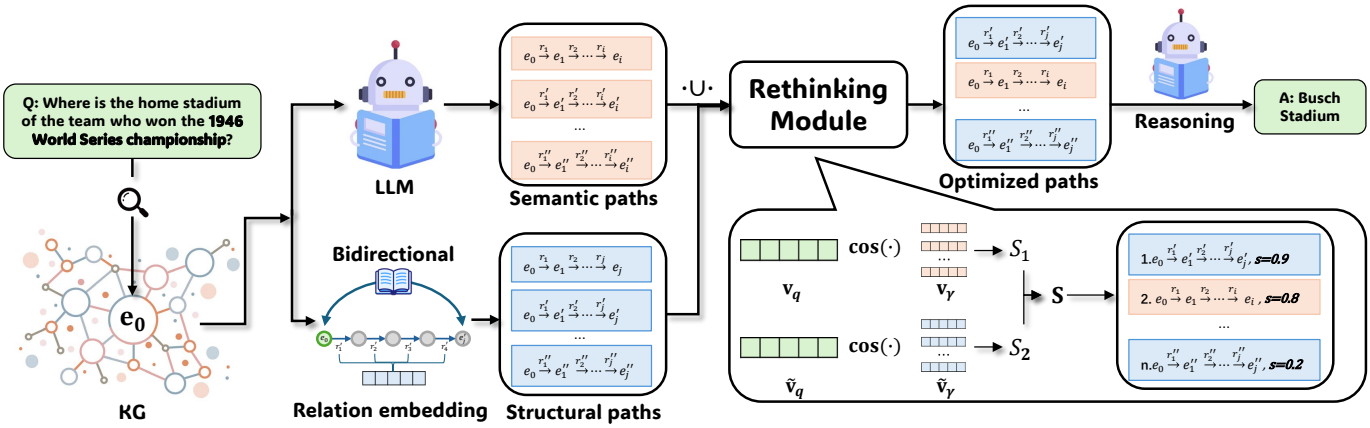


Fig. 2. 所提出方法的整体框架。给定问题和知识图谱，1）我们首先利用大型语言模型强大的语义能力生成与回答问题语义相关的推理路径。2）其次，使用第 III-A 节中描述的模块，我们从知识图谱中提取结构信息，以生成结构相关的推理路径。3）最后，我们结合这两组推理路径，使用重新思考模块进行优先排序，并将优化后的路径输入到大型语言模型中进行最终推理。

q 和一个答案 a ，我们识别出 $\gamma(q,a)$ ，即一个连接 e_q 和 e_a 的知识图谱的推理路径。然后该路径可以被视为正确回答 q 的可靠推理路径。可靠推理路径的后验分布 $\mathcal{P}(\gamma)$ 定义如下：

$$\mathcal{P}(\gamma) \simeq \mathcal{P}(\gamma \mid a,q,\mathcal{G}) = \begin{cases} \frac{1}{\Gamma}, & \exists \gamma(e_q,e_a) \in \mathcal{G}, \\ 0, & \text{else.} \end{cases} \tag{1}$$

在这里我们假设在所有可靠推理路径的集合上进行均匀先验分布，表示为 Γ ，其中 $\exists\gamma(e_q,e_a)\in\mathcal{G}$ 表示在 $\mathcal{G}$ 中存在一个推理路径实例 γ ，该路径连接了问题实体 e_q 和答案实体 e_a 。我们将生成可靠推理路径 γ 的概率定义为 $P_\alpha(\gamma \mid q)$ ，其中 α 表示 LLMs 的参数。基于这些，KL 散度可以计算为以下公式：

$$\begin{aligned} \mathcal{L}_{r_1} &= D_{\mathrm{KL}}(\mathcal{P}(\gamma \mid a,q,\mathcal{G}) \,||\, P_\alpha(\gamma \mid q)) \\ &= -\mathbb{E}_{\gamma\sim\mathcal{P}(\gamma|a,q,\mathcal{G})} \log P_\alpha(\gamma \mid q) + \text{CONST} \\ &= -\sum_{\gamma\in\Gamma^*} \mathcal{P}(\gamma \mid a,q,\mathcal{G}) \log P_\alpha(\gamma \mid q) + \text{CONST}, \\ &\simeq -\frac{1}{|\Gamma^*|} \sum_{\gamma\in\Gamma^*} \log P_\alpha(\gamma \mid q), \end{aligned} \tag{2}$$

由于推理路径 Γ 数量庞大，精确计算期望值是不可行的，因此我们通过考虑知识图谱（KG）中 e_q 和 e_a 之间的最短路径 $\gamma\in\Gamma^*\subset\Gamma$ 子集来近似期望值，并且在最终优化中省略 CONST，因为它对损失函数没有贡献。通过优化方程 2 ，我们旨在最大化生成可靠推理路径的 LLMs 的可能性。该过程有效地将知识从 KG 提炼到 LLMs 中。

A. 结构推理路径生成

按照第 III 节中描述的模块，我们利用 LLMs 强大的语义能力来提取与问题高度语义相关的推理路径。然而，我们认为仅仅依赖语义相关性对于准确推理是不够的，特别是在结构关系起关键作用的多跳推理场景中。与问题在结构上相关的推理路径可以提供正确的推理逻辑，从而实现更稳健的推理。为了应对这一限制，我们通过关系嵌入和双向分布学习增强了框架的结构挖掘能力。

我们首先将输入问题翻译成一系列推理指令。具体来说，我们使用 GloVe [28] 来获取问题 q 的初始词嵌入。这些嵌入随后通过 LSTM 编码器进行处理，最终的隐状态表示该问题。随后，利用一个递归解码器在 n 步骤后生成相应的推理指令 $\{\omega^i\}_{i=1}^n$ 。然后使用 $\{\omega^i\}_{i=1}^n$ 作为指导信号以促进实体分布的学习。为了初始化实体嵌入 $\boldsymbol{v}_e$ ，我们整合了涉及每个实体 e 的关系信息，定义如下公式：

$$\boldsymbol{v}_e^0 = \sigma\left(\sum_{\langle e,r,e'\rangle\in\mathcal{G}} \boldsymbol{v}_r \cdot W_1\right), \tag{3}$$

，其中 W_1 表示可学习的权重矩阵。与传统的实体嵌入方法相反，我们关注的是关系的嵌入，记作 $\boldsymbol{v}_r$ 。我们的理由是，为了有效地生成更准确和稳健的推理路径，重点应放在实体之间的关系上，因为它们捕捉了知识图谱中的基

Algorithm 1 结构推理路径生成

Input : Question q , Knowledge Graph $\mathcal{G}$.
Output : Reasoning Paths $\Gamma^*_{structural}$.

$\Gamma^* \leftarrow \emptyset$.;
Employ GloVe to get initial embeddings $\boldsymbol{v}_q^{initial}$ of q .
$\boldsymbol{v}_q^{initial}$ are processed through LSTM, where the final hidden state $\boldsymbol{v}_q$ represents the question.
Initialize the entity embeddings $\boldsymbol{v}_e$ by:
$\boldsymbol{v}_e^0 = \sigma\left(\sum_{\langle e,r,e'\rangle\in\mathcal{G}} \boldsymbol{v}_r \cdot W_1\right)$;
for $i \leftarrow 1$ to n **do**
  Construct a match vector $\boldsymbol{m}^i_{\langle e,r,e'\rangle}$ by:
  $\boldsymbol{m}^i_{\langle e,r,e'\rangle} = \sigma\left(\omega^i \odot W_2 \boldsymbol{v}_r\right)$;
  Aggregate matching messages from neighboring by:
  $\tilde{\boldsymbol{v}}_e^i = \sum_{\langle e,r,e'\rangle\in\mathcal{G}} P_{e'}^{i-1} \cdot \boldsymbol{m}^i_{\langle e,r,e'\rangle}$;
  Update the embeddings of entities as follows:
  $\boldsymbol{v}_e^i = \mathrm{FFN}([\boldsymbol{v}_e^{i-1}; \tilde{\boldsymbol{v}}_e^i])$;
  Probability distribution P^i could be obtained:
  $P^i = \mathrm{softmax}\left((V^i)^T \boldsymbol{W}\right)$;
  Update the loss $\mathcal{L}_{r_2}$ based on P^i .
end for
Obtain Reasoning Paths $\Gamma^*_{structural}$ based on P^n .
return Reasoning Paths $\Gamma^*_{structural}$.

本结构和依赖关系。给定一个三元组 $\langle e, r, e^{'} \rangle$ ，我们可以构建一个匹配向量 $\boldsymbol{m}^i_{\langle e,r,e' \rangle}$ 。该向量可以通过将当前推理指令 ω^i 与关系向量 $\boldsymbol{v}_r$ 进行匹配来学习：

$$\boldsymbol{m}^i_{\langle e,r,e' \rangle} = \sigma \left(\omega^i \odot W_2 \boldsymbol{v}_r \right), \tag{4}$$

其中 W_2 代表可学习的权重矩阵。接下来，我们从相邻三元组中聚合匹配信息，根据它们在上一步推理中获得的注意力分配权重：

$$\tilde{\boldsymbol{v}}^i_e = \sum_{\langle e,r,e' \rangle \in \mathcal{G}} P^{i-1}_{e'} \cdot \boldsymbol{m}^i_{\langle e,r,e' \rangle}, \tag{5}$$

其中 $P^{i-1}_{e'}$ 表示在上一步分配给实体 $e^{'}$ 的概率。此聚合有效地捕捉了知识图谱中实体的关系上下文，加强了其表示。随后，我们更新实体的嵌入，如下所示：

$$\boldsymbol{v}^i_e = \mathrm{FFN}([\boldsymbol{v}^{i-1}_e; \tilde{\boldsymbol{v}}^i_e]), \tag{6}$$

，其中 $\mathrm{FFN}(\cdot)$ 表示前馈神经网络层。该更新机制有效地将关系路径的结构编码到实体嵌入中。该模型在候选实体 P^i 上保持一个概率分布 P^i 。该分布可以形式化地定义如下：

$$P^i = \mathrm{softmax} \left((V^i)^T \boldsymbol{W} \right), \tag{7}$$

，其中 V^i 是在 i 步骤的嵌入矩阵，$\boldsymbol{W}$ 表示推导出实体分布 P^i 的可学习权重矩阵，V^i 通过方程 6 进行更新。

在训练过程中，我们认为双向分布学习在挖掘准确的结构信息和生成更全面的推理路径方面起着至关重要的作用。为了形式化这种方法，我们将从 e_q 到 e_a 的前向实体分布定义为 P^i_f ，从 e_a 到 e_q 的后向实体分布定义为 P^i_b 。这种方法的核心假设是，为了使推理过程稳定且准确，P^i_f 和 P^i_b 应该表现出高度的相似性和一致性。这个假设确保了双向分布能够相互强化，增强模块的鲁棒性。为了实现这一原则，我们引入了如下损失函数来生成结构化的推理路径：

$$\begin{aligned} \mathcal{L}_{r_2} = & D_{\mathrm{KL}} \left(P^i_f , P^*_f \right) + D_{\mathrm{KL}} \left(P^i_b , P^*_b \right) \\ & + \sum_{i=1}^{n-1} D_{\mathrm{JS}} (P^i_f, P^{n-i}_b), \end{aligned} \tag{8}$$

这里 $D_{\mathrm{JS}}(\cdot)$ 表示 Jensen-Shannon 散度，这是一种对称测度，用于量化两个概率分布之间的差异。伪代码在算法 1 中提供。

B. 重新思考推理路径

在生成推理路径方面，我们的方法与之前的工作不同，确保推理路径不仅涵盖语义相关的路径，还涵盖结构相关的路径。这种全面的对齐增强了生成推理路径的整体质量和效果。之前的方法通常是直接将所有生成的推理路径输入到 LLMs 中以获取最终答案。我们认为这种做法并不理想，因为仅仅包含正确的推理路径并不一定能从 LLMs 中得出正确答案。这种限制是由于正确的推理路径可能被放在提示的末尾或者被众多冗余的推理路径遮蔽，从而阻碍了 LLMs 有效提取和使用相关知识。因此，为了进一步提高自动推理的效率，我们认为全面的全局重新思考生成的推理路径对于提高推理准确性至关重要。

Algorithm 2 RRP 算法

Input : Question q , Knowledge Graph $\mathcal{G}$.
Parameter : λ_1 , λ_2 and θ .
Output : Reasoning Paths Γ^* , Answer a .

$\Gamma^* \leftarrow \emptyset$.;
Gnerate semantic reasoning paths Γ^*_1 ;
Gnerate structural reasoning paths Γ^*_2 ;
for $\gamma^i \in (\Gamma^*_1 \cup \Gamma^*_2)$ **do**
  Compute the semantic similarity score $S_1(q, \gamma^i)$;
  Compute the structural similarity score $S_2(q, \gamma^i)$;
  Combine these socre by:
  $\boldsymbol{S}(q, \gamma^i) = \lambda_1 \cdot S_1(q, \gamma^i) + \lambda_2 \cdot S_2(q, \gamma^i)$;
  if $\boldsymbol{S}(q, \gamma^i) > \theta$ **then**
    Γ^* .append (γ^i) ;
  else
    Filter γ^i ;
  end if
end for
Sort (Γ^*) by the $\boldsymbol{S}$;
Feed (q, Γ^*) to LLMs to get answer a ;
return answer a .

根据 Zellig Harris 的分布假设，出现在相似语境中的语言单位往往具有相似的意义。在 LLM 的语境中，问题及其对应的推理路径通常出现在训练语料库的相同或相似语境中。因此，我们假设在 LLM 的特征空间中，问题及其正确推理路径的嵌入是相似的。具体而言，我们利用第 III 节中描述的模块来获得问题的嵌入 $\boldsymbol{v}_q$ 以及第 i 个推理路径的嵌入 $\boldsymbol{v}^i_\gamma$ 。然后定义问题与推理路径之间的语义相似度评分为以下公式：

$$S_1(q, \gamma^i) = \cos \left(\boldsymbol{v}_q, \boldsymbol{v}^i_\gamma \right), \tag{9}$$

，其中 $\cos(\cdot)$ 表示两个向量之间的余弦相似性。此外，我们断定问题与推理路径之间的结构相似性对于生成准确答案至关重要。如在第 III-A 节详细描述的那样，我们的方法主要在编码问题时利用关系信息。在训练过程中，我们以双向方式学习问题实体和答案实体之间的有效推理路径。这种双向学习固有地将有效推理路径的结构信息嵌入到问题表现中。因此，我们使用以下公式量化问题与推理路径之间的结构相似性评分：

$$S_2(q, \gamma^i) = \cos \left(\tilde{\boldsymbol{v}}_q, \frac{1}{n} \sum_{j=1}^{n} \tilde{\boldsymbol{v}}^{i^j}_\gamma \right), \tag{10}$$

，其中 $\tilde{\boldsymbol{v}}$ 表示从第 III-A 节中的模块获得的嵌入。由于推理路径由多个实体组成，我们直观地通过平均其构成实体的嵌入来表示路径的嵌入。随后，我们结合语义相似性评分和结构相似性评分来量化每个推理路径在准确回答问题中的总体重要性。这些评分的整合计算如下：

$$\boldsymbol{S}(q, \gamma^i) = \lambda_1 \cdot S_1(q, \gamma^i) + \lambda_2 \cdot S_2(q, \gamma^i), \tag{11}$$

其中 λ_1 和 λ_2 是超参数。利用计算出的重要性分数 $\boldsymbol{S}(q, \gamma^i)$ ，我们按照重要性降序排列与问题 q 对应的推理路径。然后我们设定一个阈值 θ ，以消除与 q 不足够相关的推理路径。具体来说，我们消除那些重要性分数低于 θ 的路径。在为大语言模型准备问答输入时，我们将剩余的推理

Category	Method	WebQSP		CWQ	
		Hits@1	F1	Hits@1	F1
Embedding-based	KV-Mem [31]	46.7	34.5	18.4	15.7
	EmbedKGQA [32]	66.6	-	45.9	-
	NSM [33]	68.7	62.8	47.6	42.4
	TransferNet [34]	71.4	-	48.6	-
	KGT5 [35]	56.1	-	36.5	-
Retrieval-based	GraftNet [36]	66.4	60.4	36.8	32.7
	PullNet [37]	68.1	-	45.9	-
	SR+NSM [38]	68.9	64.1	50.2	47.1
	SR+NSM+E2E [38]	69.5	64.1	49.3	46.3
LLMs	Flan-T5-xl [39]	31.0	19.9	14.7	13.2
	Alpaca-7B [2]	51.8	34.3	27.4	22.2
	LLaMA2-Chat-7B [1]	64.4	28.0	34.6	16.9
	ChatGPT	66.8	39.3	39.9	28.5
	ChatGPT+CoT [40]	75.6	-	48.9	-
LLMs+KGs	KD-CoT [26]	68.6	52.5	55.7	-
	UniKGQA [19]	77.2	72.2	51.2	49.1
	ToG+ChatGPT [16]	76.2	-	58.9	-
	ToG+LLaMA2-70B [16]	68.9	-	57.6	-
	KG-CoT [18]	84.9	-	62.3	-
	RoG [17]	85.7	70.8	62.6	56.2
	SubgraphRAG+ChatGPT [27]	83.1	69.2	56.3	49.1
	SubgraphRAG+LLaMA3-8B [27]	86.6	70.6	57.0	47.2
	RRP(ours)	90.0	72.5	64.5	56.5

TABLE I
在两个标准数据集上与各种基准的性能比较。

路径按重要性递减的顺序包括在提示中。此方法确保大语言模型专注于最相关的推理路径，从而提高生成答案的准确性。对于大语言模型的微调，我们遵循标准的提示调优过程 [17] 。为了清晰直观地表达 RRP 算法，伪代码在算法 2 中提供。该算法利用第 III 节和第 III-A 节中介绍的模块来生成全面而准确的推理路径。然后通过重新思考过程对这些路径进行优化，包括根据重要性重新排序、消除冗余路径，并将优化后的路径输入到大语言模型中以增强推理能力。

我们在两个广泛使用的基准数据集上评估了该方法的有效性。我们遵循之前工作的设置，使用相同的训练和测试划分比例以进行公平比较。

WebQuestionsSP (WebQSP) [29] 包含 4737 个自然语言问题，每个问题都与 Freebase 中的可靠推理路径相关联。这个数据集关注需要简单推理路径的问题。

ComplexWebQuestions (CWQ) [30] 包含 34699 个复杂的自然语言问题。与主要集中于简单事实查询的 WebQSP 不同，CWQ 强调多跳推理，因为可靠的推理路径通常涉及穿越知识图中的多个关系。

对于第 III 节中描述的模块，我们利用 LLaMA2-Chat-7B 作为骨干模型。训练过程配置了以下超参数：训练的轮数设置为 3，批量大小设置为 4，学习率设置为 $2e^{-5}$ 。对于第 III-A 节中详细说明的模块，训练在 80 个轮次下进行，批量大小为 40，学习率为 $4e^{-4}$ 。重新思考模块的超参数在两个数据集上是不同的：对于 WebQSP，参数 (λ_1, λ_2) 设置为 $(0.5, 0.5)$ ，而对于 CWQ，它们设置为 $(0.1, 0.9)$ 。根据以往的研究，我们采用 Hits@1 和 F1 分数作为主要评估指标来评估我们方法的有效性。Hits@1 评估预测答案中的前 1 名与真实答案匹配的问题的比例。F1 分数则考虑到一个问题对应多个正确答案的情况。它考虑了精确率和召回率之间的权衡，从而提供了模型覆盖率和准确性的综合测量。

C. 基线

我们通过将所提出的方法与代表各种不同类别的强基线进行比较来评估它。这些基线分为四个不同的类型：

基于嵌入的方法：这些方法侧重于构建高质量的实体和关系的向量表示，以促进准确的推理和问答。

- KV-Mem [31] 是一种经过修改的记忆网络版本，它通过在记忆读取操作的寻址和输出阶段使用不同的编码来提高文档的可读性。
- EmbedKGQA [32] 引入了使用知识图嵌入来进行问答，放松了对答案选择必须在预先指定的局部邻域内进行的要求。
- NSM [33] 采用了一个师生框架，生成更可靠的中间监督信号，增强了推理性能。
- TransferNet [34] 通过在每个步骤动态关注问题的不同部分，实现了多步骤推理。它以可微分的方式计算关系的激活得分，并在激活的关系上转移实体得分。
- KGT5 [35] 表明，现成的编码器-解码器 Transformer 模型可以作为一个可扩展和多功能的知识图谱嵌入（KGE）模型运行，在不完整的知识图谱问答任务中取得良好的效果。

基于检索的方法：通过整合外部检索机制来增强问答过程的方法。

- GraftNet [36] 引入了一种受个性化 PageRank 启发的定向传播方法，该方法通常用于信息检索。
- PULLNet [37] 采用迭代过程构建特定问题的子图，以获取与该问题相关的信息。在每次迭代中，GCN 通过对语料库和/或知识库进行检索操作来识别需要扩展的子图节点。一旦子图构建完成，一个类似的 GCN 从子图中提取最终答案。

- SR [38] 是一个可训练的子图检索器，独立于随后的推理过程进行操作。这种解耦可以形成一个即插即用的框架，以增强任何面向子图的模型。

大型语言模型（LLMs）：[1], [2], [39], [40] 利用预训练的大规模神经语言模型，根据其广泛的语言理解能力生成答案的方法。

LLMs+KGs 方法：一种混合方法，将 LLMs 的解释能力与 KGs 中的信息相结合，以提高推理准确性。

- KD-CoT [26] 引入了一种知识驱动的思维链框架，通过与外部知识的交互验证和修改思维链中的推理轨迹。该方法缓解了幻觉和错误传播，提高了推理准确性。
- UniKGQA [19] 提出了一种基于预训练语言模型的方法，该方法在模型架构和参数学习中统一了检索和推理，简化了问答过程。
- ToG [16] 引入了一种新颖的范式，其中 LLM 代理在知识图谱 (KG) 上迭代地执行束搜索，有效地发现和探索推理路径。
- KG-CoT [18] 在思维链框架的基础上，通过引入一个小规模的、逐步的图推理模型来促进对知识图的推理。
- RoG [17] 提出了一个规划-检索-推理框架。RoG 生成关系路径，然后用这些路径从知识图谱中检索推理路径。这些检索到的路径作为输入供 LLMs 进行更具结构化和准确的推理。
- SubgraphRAG [27] 提出了一个创新框架，该框架将轻量级多层感知机与并行三重打分机制相结合，用于高效且灵活的子图检索，同时编码方向性结构距离以增强检索效果。

如表 I 所示，我们首先按类别分析结果。对于依赖嵌入的传统方法和依赖检索的方法，其性能总体保持稳定，但表现出明显的瓶颈。当单独使用 LLMs 进行推理时，其性能不如一些基于嵌入和检索的方法。这是因为虽然 LLMs 具有强大的语义能力，但缺乏有效的结构化挖掘。然而，将 LLMs 与 KGs 结合的方法整体上优于其他类别，因为 KGs 提供的结构化知识弥补了 LLMs 在挖掘结构信息方面的不足。这些结果表明，将 LLMs 与 KGs 结合构成了一种有效的推理任务范式。此外，与 LLMs+KGs 类别中的其他同类方法相比，我们提出的方法在 WebQSP 和 CWQ 数据集上实现了最先进的性能。

具体来说，在 WebQSP 数据集上，我们的方法在 Hits@1 上超过了最近的 RoG 方法 4.3 个百分点，并在 F1 评分上取得了 1.7 个百分点的提升。这种改进可以归因于我们方法中包含了更全面的推理路径和推理路径的重新思考机制。此外，我们的方法超过了其他类别方法 15 到 60 个百分点，进一步证明了结合大型语言模型和知识图谱的有效性。同样地，在 CWQ 数据集上，我们的方法在 Hits@1 上比 RoG 高出 2.0 个百分点，证明了其在处理复杂查询时的有效性。来自其他类别的方法在这些具有挑战性的查询上举步维艰，因为此类查询需要高语义理解和深层结构理解。值得注意的是，这些结果是使用只有 7B 参数的 LLM 实现的，这表明我们的方法无需依赖过于庞大的模型规模即可提供具有竞争力的性能，使其成为资源受限场景中的实用解决方案。

D. 即插即用研究

由于我们的方法在推理过程中与任何大型语言模型兼容，我们评估其集成在各种模型中的推理性能提高了多少。为了进行这个评估，我们首先使用我们的框架为每个问题生成一组高质量的推理路径。然后，我们将原始问题和相应的推理路径提供给不同的 LLM，使它们能够直接基于这些路径推导出答案，而无需额外的微调。表 II 总结了这次实验的结果。在 WebQSP 数据集中，集成我们的方法为所有测试模型的 Hits@1 带来了显著的提升。例如，Flan-T5 xl 相比没有使用我们的推理路径的基线性能，增加了 38.71%。其他 LLM 在 Hits@1 中也超过了 20% 的改进。这些结果证实了提供结构化的推理链显著增强了模型定位正确答案的能力，即使基础的语言模型保持不变。在 CWQ 数据集中，类似的趋势出现：GPT-4o mini 在 Hits@1 中表现出 30.25% 的增长，其他所有测试的 LLM 显示出超过 20% 的提升。这些数据集中的一致增强明确表明，我们的方法在不同的模型架构下有效地解锁了更强的推理能力。

Methods	WebQSP		CWQ	
	Hits@1	Recall	Hits@1	Recall
GPT-4o-mini	67.08	40.64	41.43	35.95
GPT-4o-mini + RRP	91.10	81.52	71.68	66.79
ChatGPT	66.77	49.27	39.90	35.07
ChatGPT + RRP	89.93	79.34	63.07	57.99
Alpaca-7B	51.78	33.65	27.44	23.62
Alpaca-7B + RRP	78.13	59.34	48.40	42.00
LLaMA2-Chat-7B	64.37	44.61	34.60	29.91
LLaMA2-Chat-7B + RRP	86.79	75.39	59.78	54.76
Flan-T5-xl	30.95	17.08	14.69	12.25
Flan-T5-xl + RRP	69.66	45.90	40.61	34.27

TABLE II
所提出方法在不同 LLMs 上的性能提升。

特别值得注意的是，GPT-4o mini 结合我们的推理路径生成框架，在 WebQSP 和 CWQ 上提供了最新的准确率。这一结果突出了两个重要点。首先，我们的方法生成的推理路径提供了关键的上下文，单靠大型语言模型可能无法从查询文本中可靠推断出。其次，由于不需要微调，我们的方法为在资源受限的情况下提供了一种实用的方法来提升性能，此时重新训练或调整大型模型是不可行的。总之，这项即插即用的研究表明，我们的推理路径生成模块可以在各种大型语言模型中作为通用且有效的增强工具。它在复杂的问答任务中产生了显著且一致的改进。

E. 超参数敏感性分析

我们提出的模型包括两个超参数，θ 和 λ 。参数 θ 表示用于过滤冗余推理路径的阈值；较低的阈值会保留较多的路径。参数 λ 控制语义信息与结构信息之间的平衡。具体来说，λ_1 代表语义信息的权重，$\lambda_2 = 1 - \lambda_1$ 代表结构信息的权重，且它们共同满足 $\lambda_1 + \lambda_2 = 1$。为了彻底评估我们方法对各种超参数的敏感性，我们进行了全面的实验系列。

图 3 展示了改变 θ 的效果。当 θ 从最小值增加到 0.6 时，模型性能逐渐提高，并在 $\theta = 0.6$ 时达到最佳状态。超过这一点，θ 的进一步增加会导致性能下降，尤其是在 $\theta = 0.8$ 时观察到明显的下降。这种行为可以解释为：增加 θ 更积极地过滤掉不相关的推理路径，从而减少它们对正确推理的负面影响。在 $\theta = 0.6$ 时，模型达到了最佳平衡，有效地去除了绝大多数不相关路径，同时保留了大部分有效路径。然而，如果过滤阈值过高，大量推理路径会被丢弃，其中包括一些有效的路径，导致性能下降。

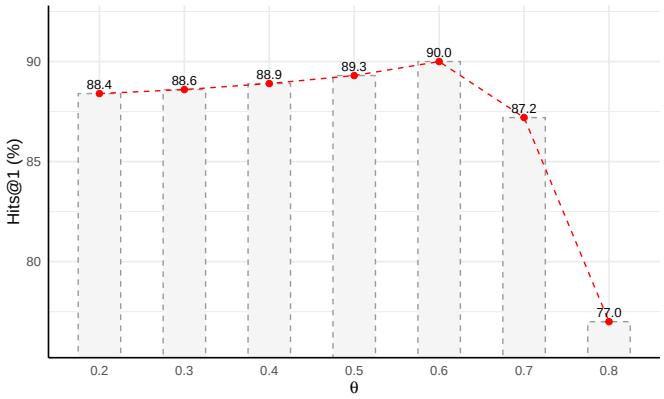


Fig. 3. 在 WebQSP 上不同 θ 值的性能。

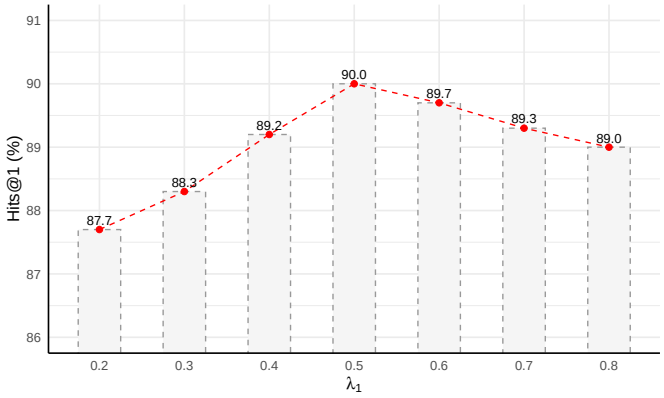


Fig. 4. 在 WebQSP 上不同 λ_1 值的表现。

图 4 显示了随着 λ_1（以及从而 λ_2）的变化而产生的结果。随着 λ_1 从其最小值增加到 0.5，模型性能再次提升，并在 $\lambda_1 = 0.5$ 时达到峰值。λ_1 的进一步增加导致性能下降。这一趋势源于对于 WebQSP 数据集而言，语义信息和结构信息同等重要。当 λ_1 增加时，模型会赋予语义更多的权重，这最初提升了性能；在 $\lambda_1 = 0.5$ 时，语义和结构信息的权重达到了完美的平衡。超过 $\lambda_1 = 0.5$ 后，结构信息的权重过轻，这对模型利用结构线索的能力产生了负面影响，从而导致整体性能下降。

F. 对不同大小知识图谱的鲁棒性

Characteristics	Knowledge Graph	
	Freebase subgraph	Wiki-Movie
Number of entities	2,566,291	43,234
Number of relations	7,058	9
Number of triples	8,309,195	133,582
Accuracy	Medium	High
Average inference time	2.9s	2.6s

TABLE III
用于验证模型鲁棒性的知识图谱比较。

在主要结果中，我们首先在 WebQSP 和 CWQ 上展示了所提方法的有效性，这两个数据集都是建立在大规模的 Freebase 知识图谱之上的。然而，由于现实世界中应用的知识图谱大小可能有很大差异，我们进一步通过引入不同规模的额外知识图谱进行比较，以评估我们方法的鲁棒性。

如表 III 所示，Freebase 包含 2,566,291 个实体，7,058 个关系，总共有 8,309,195 个三元组，作为一个具有代表性的大规模知识图谱。由于它跨越多个领域并包含大量信息，其整体数据质量评估为中等。总而言之，Freebase 既大规模又高度复杂。相比之下，Wiki-Movie 包含 43,234 个实体，9 个关系和 133,582 个三元组，代表中等规模知识图谱。由于 Wiki-Movie 限定于单一领域并以高精度构建，其数据质量被归类为高。总体而言，Wiki-Movie 规模适中，与 Freebase 相比，呈现出较低的结构复杂性。

Methods	WebQSP(Freebase)	MetaQA-3(Wiki-Movie)
	Hits@1	
RRP(ours)	90.0	99.5
RoG	85.7	89.0
GraftNet	66.4	77.7
PullNet	68.1	91.4
EmbedKGQA	66.6	94.8

TABLE IV
所提出方法对不同知识图谱的鲁棒性。

对于比较实验，我们选择了一组具有代表性的基线方法；结果在表 IV 中报告，其中 WebQSP 对应 Freebase 评价，MetaQA-3 对应 Wiki-Movie 评价。在大规模和中等规模知识图谱中，我们提出的方法在性能上超越所有基线，取得了显著的优势。这个结果进一步验证了 RRP 在现实环境中的鲁棒性。此外，我们计算了我们方法在两个知识图谱上的平均推理时间，发现它们之间的差异可以忽略不计。这个观察进一步确认了 RRP 框架始终保持稳健的计算效率，展示了其可扩展性和稳定性能，不论知识图谱的规模如何。

G. 消融研究

Semantic	Structural	Rethinking	Hits@1	F1
✓	✓	✓	90.0	72.5
✓	✓	✗	88.3	71.2
✓	✗	✓	86.7	69.5
✗	✓	✓	77.4	67.9
✓	✗	✗	84.8	68.6
✗	✓	✗	74.2	67.3

TABLE V
对所提出方法在 WebQSP 上的消融研究。

如表 V 所示，我们进行了消融研究以评估我们提出的框架中每个组件的贡献。结果表明，语义推理路径生成模块本身表现强劲，这归功于 LLMs 的高级推理能力。相较之下，结构推理路径生成模块被设计用来通过遍历知识图谱中的显式连接来探索多跳路径。通过跟随图谱边缘，它发现将查询实体链接到候选答案的关系链。然而，这种专注于结构性探索的方法可能导致它忽视那些只涉及一跳或两跳的简单直接的推理路径。换句话说，它可能会错过那些语义丰富但从纯结构启发式来看并不明显的短路径。

当这两个模块结合时，其优点相辅相成。语义组件带来了灵活性和更广泛的覆盖面，而结构组件则确保了重要的多跳连接被考虑。因此，整合后的模型可以访问更全面和准确的推理路径集，涵盖直接推理和更长的多步骤推理。此外，加入重新思考模块在与前面组件中的任一个或两者结合使用时进一步提升了性能。该模块对候选推理路径进行第二次遍历，对其进行评分和排名以识别最相关的链条。通过对有助于得出最终答案的路径赋予更高的权重，并减

Question	The newspaper Zerkalo Nedeli is circulated in an area that has what as the official language?
Reasoning Paths of RoG	Zerkalo Nedeli → book.periodical.language → English Language Zerkalo Nedeli → book.periodical.language → Russian Language Zerkalo Nedeli → book.periodical.language → Ukrainian Language
Reasoning Paths of RRP	Reasoning paths are arranged in descending order of significance: Path 1(structural): Zerkalo Nedeli → periodicals.newspaper_circulation_area.newspapers → Ukraine → location.country.languages_spoken → Ukrainian Language Path 2 (semantic): Zerkalo Nedeli → book.periodical.language → Ukrainian Language Path 3 (semantic): Zerkalo Nedeli → book.periodical.language → Russian Language Path 4 (semantic): Zerkalo Nedeli → book.periodical.language → English Language
Outputs	RoG: Based on the reasoning paths provided, the official language of the area where Zerkalo Nedeli is circulated is 俄语 RRP: Based on the reasoning paths provided in order of significance, the official language of the area where Zerkalo Nedeli is circulated is 乌克兰语
Ground Truth	Ukrainian Language.

TABLE VI
RRP 案例研究。RRP 可以生成更全面的推理路径，并根据重要性重新思考它们，从而使 LLMs 获得更有效的指导。

少冗余或不太相关路径的影响，重新思考模块减轻了噪声，并提升了整体推理的准确性。

H. 案例研究

如表 VI 所示，该案例研究体现了 RRP 的优越性。首先，对于这个问题，SOTA 方法仅产生“镜报”的“周期语言”。相比之下，RRP 成功生成了报纸流通地区的语言，这与问题的背景更相关。这表明单靠 LLMs 可能不足以准确理解这类问题中复杂的知识要求及其固有的逻辑相互依赖关系。RRP 有效地揭示了连接相关知识的逻辑链，生成更全面和准确的推理路径。其次，基于这些推理路径，SOTA 方法提供了错误的答案，而 RRP 通过根据路径的重要性进行优先排序，准确地解决了这一问题。总体而言，由 RRP 框架生成的全面且逻辑有序的推理路径，为 LLM 推理提供了精炼和针对性的指导。

I. 专家设计的提示模板

语义推理路径生成模块旨在生成有效的推理路径，这可以作为回答问题的指导。提示模板如图 5 所示，其中 < 问题 > 表示问题的内容。

Semantic Paths Generation Prompt

Please generate the valid reasoning paths that can be helpful for answering the following question:

<Question>

Fig. 5. 用于生成语义路径的提示模板。

在我们使用 RRP 框架得到全面且有序的推理路径后，LLM 推理过程作为最终步骤。通过输入推理路径和问题，引导 LLM 生成正确答案。有关更多详细信息，提示模板如图 6 所示，其中 < 推理路径 > 表示有序的推理路径实例。

IV. 结论

在本文中，我们提出了 RRP，这是一种利用知识图谱增强大型语言模型推理的新方法。我们的框架战略性地结合了 LLM 的语义优势和通过关系嵌入及双向分布学习获得的丰富结构信息，确保推理过程的广度和精确性。为了进一步提高推理质量，我们引入了一个重思考模块，系统地组织、评估和过滤候选推理路径，使得能够识别出最连贯和有影响力的路径用于下游的 LLM 推理。在两个公共数据集上进行的广泛实验表明，RRP 始终达到当前最先进的性能，超越了多种竞争基线。重要的是，即使在仅使用 7 亿参数的相对轻量级 LLM 时，我们也展示了这些收益，这强调了我们方法的效率和可扩展性。除经验有效性外，RRP 的模块化、即插即用设计提供了与任何现有 LLM 架构的无缝兼容性，便于轻松整合到多种现实应用中，而无需进行大量再训练或模型修改。展望未来，RRP 为未来在知识支撑的语言理解方面的进步奠定了坚实的基础。其针对特定查询提炼高质量推理路径的能力不仅减轻了幻觉现象并提升了事实正确性，还为更透明和可解释的 AI 开辟了新的途径。我们真诚地期望，我们框架的核心想法将带来见解并激励后续关于复杂知识图谱增强技术和更稳健推理范式的研究，在不断发展的大规模语言智能领域中。

LLM Reasoning Prompt

Instructions:
Please use the reasoning paths provided below to answer the question. The reasoning paths are listed in order of importance, with the first being the most important. Your task is to derive the simplest possible answer and return all potential answers as a list.

Reasoning Paths:
<Reasoning Paths>

Question:
<Question>

Fig. 6. 大型语言模型推理的提示模板。

References

[1] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.

[2] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, “Stanford alpaca: An instruction-following llama model,” https://github.com/tatsu-lab/stanford_alpaca, 2023.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-1144.html

[5] Q. Zhang, S. Chen, Y. Bei, Z. Yuan, H. Zhou, Z. Hong, J. Dong, H. Chen, Y. Chang, and X. Huang, "A survey of graph retrieval-augmented generation for customized large language models," *arXiv preprint arXiv:2501.13958*, 2025.

[6] Q. Zhang, J. Dong, H. Chen, D. Zha, Z. Yu, and X. Huang, "Knowgpt: Knowledge graph based prompting for large language models," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 6052–6080. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/0b8705a611ed1ce19cdb759031078705-Paper-Conference.pdf

[7] T. Korbak, H. Elsahar, G. Kruszewski, and M. Dymetman, "Controlling conditional language models without catastrophic forgetting," in *International Conference on Machine Learning*. PMLR, 2022, pp. 11 499–11 528.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models." in *The Tenth International Conference on Learning Representations*. OpenReview.net, 2022. [Online]. Available: http://dblp.uni-trier.de/db/conf/iclr/iclr2022.html#HuSWALWWC22

[9] Z. Gekhman, G. Yona, R. Aharoni, M. Eyal, A. Feder, R. Reichart, and J. Herzig, "Does fine-tuning LLMs on new knowledge encourage hallucinations?" in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 7765–7784. [Online]. Available: https://aclanthology.org/2024.emnlp-main.444/

[10] J. He, C. Zhou, X. Ma, and et al., "Towards a unified view of parameter-efficient transfer learning," in *The Tenth International Conference on Learning Representations*, 2022.

[11] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2312.10997

[12] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning, "RAPTOR: Recursive abstractive processing for tree-organized retrieval," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=GN921JHCRw

[13] Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, "Lightrag: Simple and fast retrieval-augmented generation," 2024.

[14] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," 2025. [Online]. Available: https://arxiv.org/abs/2404.16130

[15] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1842–1850.

[16] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. Ni, H.-Y. Shum, and J. Guo, "Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=nnVO1PvbTv

[17] L. LUO, Y.-F. Li, R. Haf, and S. Pan, "Reasoning on graphs: Faithful and interpretable large language model reasoning," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=ZGNWW7xZ6Q

[18] R. Zhao, F. Zhao, L. Wang, X. Wang, and G. Xu, "Kg-cot: Chain-of-thought prompting of large language models over knowledge graphs for knowledge-aware question answering," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 6642–6650, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2024/734

[19] J. Jiang, K. Zhou, X. Zhao, and J.-R. Wen, "UniKGQA: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=Z63RvyAZ2Vh

[20] Z. Yang, Q. Liu, T. Pang, H. Wang, H. Feng, M. Zhu, and W. Chen, "Self-distillation bridges distribution gap in language model fine-tuning," *arXiv preprint arXiv:2402.13669*, 2024.

[21] Z. Gekhman, G. Yona, R. Aharoni, M. Eyal, A. Feder, R. Reichart, and J. Herzig, "Does fine-tuning llms on new knowledge encourage hallucinations?" *arXiv preprint arXiv:2405.05904*, 2024.

[22] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[23] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua, "Retrieving and reading: A comprehensive survey on open-domain question answering," *arXiv preprint arXiv:2101.00774*, 2021.

[24] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu *et al.*, "Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv:2107.02137*, 2021.

[25] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, "K-bert: Enabling language representation with knowledge graph," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 2901–2908.

[26] K. Wang, F. Duan, S. Wang, P. Li, Y. Xian, C. Yin, W. Rong, and Z. Xiong, "Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering," *arXiv preprint arXiv:2308.13259*, 2023.

[27] M. Li, S. Miao, and P. Li, "Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation," in *International Conference on Learning Representations*, 2025.

[28] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[29] W.-t. Yih, M. Richardson, C. Meek, M.-W. Chang, and J. Suh, "The value of semantic parse labeling for knowledge base question answering," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 201–206. [Online]. Available: https://aclanthology.org/P16-2033

[30] A. Talmor and J. Berant, "The web as a knowledge-base for answering complex questions," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 641–651. [Online]. Available: https://aclanthology.org/N18-1059

[31] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1400–1409. [Online]. Available: https://aclanthology.org/D16-1147

[32] A. Saxena, A. Tripathi, and P. Talukdar, "Improving multi-hop question answering over knowledge graphs using knowledge base embeddings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 4498–4507. [Online]. Available: https://aclanthology.org/2020.acl-main.412

[33] G. He, Y. Lan, J. Jiang, W. X. Zhao, and J.-R. Wen, "Improving multi-hop knowledge base question answering by learning intermediate supervision signals," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, ser. WSDM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 553–561. [Online]. Available: https://doi.org/10.1145/3437963.3441753

[34] J. Shi, S. Cao, L. Hou, J. Li, and H. Zhang, "TransferNet: An effective and transparent framework for multi-hop question answering over relation graph," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4149–4158. [Online]. Available: https://aclanthology.org/2021.emnlp-main.341

[35] A. Saxena, A. Kochsiek, and R. Gemulla, "Sequence-to-sequence knowledge graph completion and question answering," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2814–2828. [Online]. Available: https://aclanthology.org/2022.acl-long.201

[36] H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, and W. Cohen, "Open domain question answering using early fusion of knowledge bases and text," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4231–4242. [Online]. Available: https://aclanthology.org/D18-1455

[37] H. Sun, T. Bedrax-Weiss, and W. Cohen, "PullNet: Open domain question answering with iterative retrieval on knowledge bases and text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2380–2390. [Online]. Available: https://aclanthology.org/D19-1242

[38] J. Zhang, X. Zhang, J. Yu, J. Tang, J. Tang, C. Li, and H. Chen, "Subgraph retrieval enhanced model for multi-hop knowledge base question answering," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5773–5784. [Online]. Available: https://aclanthology.org/2022.acl-long.396

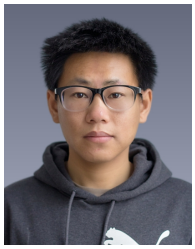
[39] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024. [Online]. Available: http://jmlr.org/papers/v25/23-0870.html

[40] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

Yilin Xiao is a first-year Ph.D. candidate at the Department of Computing, The Hong Kong Polytechnic University. Before that, he received a bachelor's degree in Dalian University of Technology and a master's degree in Wuhan University. His research interests include KGs, reasoning, LLMs, and RAG. His works have been published in venues such as ACL, ACM Multimedia, IEEE TIV, IEEE TGRS and etc. He also serves as a reviewer for ACM Multimedia, IPM and etc.

Chuang Zhou is a third-year Ph.D. candidate at the Department of Computing, The Hong Kong Polytechnic University. Before that, he received a bachelor's degree in Economics and Statistics from University College London and a master's degree in Statistics from The London School of Economics and Political Science. He is currently a member of the DEEP Lab supervised by Dr. Xiao Huang. His research interests include LLMs, recommendation, and GraphRAG. He has published several papers, including ACL, EMNLP, COLING and etc.

Qinggang Zhang is a fourth-year Ph.D. candidate at the Department of Computing, The Hong Kong Polytechnic University. Before that, he received a bachelor's degree in Engineering in Computer Science from Northwestern Polytechnical University. He is currently a member in DEEP Lab supervised by Dr. Xiao Huang. His research interests include KGs, LLMs, RAG and Text-toSQL. He has published over 20 papers while serving as reviewers for NeurIPS, ICML, ICLR, KDD, IEEE TKDE and IEEE TPAMI.



Bo Li is an assistant professor in the Department of Computing at the Hong Kong Polytechnic University. He received his Ph.D. in Computer Science from Stony Brook University. He is broadly interested in algorithms, AI, and computational economics, including problems related to resource allocation, game theory, online algorithms, and their applications to Blockchain and machine learning.

Qing Li (Fellow, IEEE) received the BEng degree from Hunan University, Changsha, China, and the MSc and PhD degrees from the University of Southern California, Los Angeles, all in computer science. He is currently a chair professor (Data science) and the head of the Department of Computing, Hong Kong Polytechnic University. He is a fellow of IET, a member of ACM SIGMOD and IEEE Technical Committee on Data Engineering. His research interests include object modeling, multimedia databases, social media, and recommender systems. He has been actively involved in the research community by serving as an associate editor and reviewer for technical journals, and as an organizer/co-organizer of numerous international conferences. He is the chairperson of the Hong Kong Web Society, and also served/is serving as an executive committee (EXCO) member of IEEE-Hong Kong Computer Chapter and ACM Hong Kong Chapter. In addition, he serves as a councilor of the Database Society of Chinese Computer Federation (CCF), a member of the Big Data Expert Committee of CCF, and is a Steering Committee member of DASFAA, ER, ICWL, UMEDIA, and WISE Society.

Xiao Huang is an Assistant Professor in the Department of Computing at The Hong Kong Polytechnic University. He earned his Ph.D. in Computer Engineering from Texas A & M University in 2020, an M.S. in Electrical Engineering from the Illinois Institute of Technology in 2015, and a B.S. in Engineering from Shanghai Jiao Tong University in 2012. His scholarly contributions are highly regarded within the academic community, amassing over 4,900 citations. He received the Best Paper Award Honorable Mention at SIGIR 2023. He has led or completed seven research projects as Principal Investigator, securing funding exceeding 5 million HKD. He serves as a PhD Symposium Chair for ICDE 2025.