Progressive Planning and Reinforced Reasoning: Large Language Model-Guided Multi-hop Question Answering over Knowledge Graph

Xiang Li[1], **Runhai Jiao**[1]*, **Ruifan Li**[2], **Dongnan Wu**[1], **Ruojiao Qiao**[1], **Lei Liu**[1]
[1]School of Control and Computer Engineering, North China Electric Power University, China
[2]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
{120212227306,runhaijiao,120221080413,120242227168,120242227112}@ncepu.edu.cn,
rfli@bupt.edu.cn

Abstract

Reinforcement learning, with its interpretable path reasoning, has emerged as a promising paradigm for multi-hop question answering over knowledge graphs. However, existing approaches suffer from two inherent limitations: (1) lacking effective intermediate guidance, agents often fall into aimless exploration when confronted with complex multi-hop questions; and (2) policy networks focus on local neighborhood information, making it difficult to anticipate the long-term consequences of decisions. To address these challenges, we propose a **P**rogressive **P**lanning and **R**einforced **R**easoning (PPRR) framework. Specifically, we introduce large language models as multi-hop reasoning planners, converting decomposed sub-question sequences into stepwise decision guidance and thereby granting the agent human-like, step-by-step problem-solving capabilities. In addition, we design a structure-aware lookahead policy network, which explicitly models inter-node dependencies along the multi-hop reasoning process and performs lookahead value evaluations for candidate actions, thereby enhancing the agent's global state awareness and decision foresight in complex environments. Finally, we conducted extensive experiments on four public multi-hop question answering benchmarks and one domain-specific dataset. The results demonstrate that our framework surpasses state-of-the-art methods while demonstrating strong generalization.

1 Introduction

Knowledge graphs (KGs), as the primary carriers of structured knowledge, have become essential infrastructure for intelligent question answering (QA) (Saxena et al., 2020; Ji et al., 2024; Shen et al., 2025), information retrieval(Xu et al., 2024b; Li et al., 2025), recommendation (Jiang et al., 2024; Patel, 2025) and beyond. Compared with traditional QA systems, knowledge graph question answering (KGQA) can directly leverage the relations among entities, thereby exhibiting stronger reasoning capability and interpretability (Bi et al., 2025; Martynova et al., 2025; Wang et al., 2025a). Within KGQA, multi-hop question answering (MHQA) is particularly challenging: it requires the model to aggregate evidence across multiple triples, and has attracted considerable research interest (Liu et al., 2025; Chen et al., 2025; Xu et al., 2025; Zhou et al., 2025).

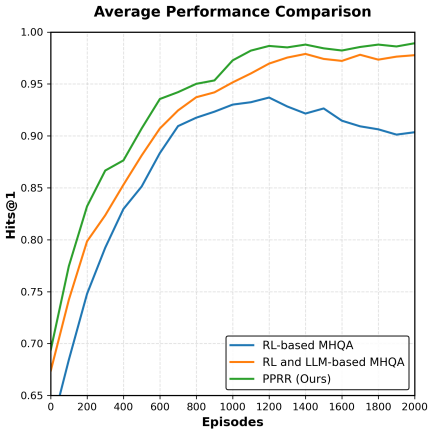

Figure 1: Average Performance comparison on the PQL-3H dataset. Curves show average Hits@1 scores on the validation set across 15 training runs for three multi-hop question answering methods: RL-based MHQA, RL and LLM-based MHQA, and our PPRR framework.

Given the sequential decision nature of multi-hop reasoning, the reinforcement learning (RL) paradigm has been widely adopted for MHQA (Wu et al., 2025; Zhang et al., 2025a). The core idea is to cast the task as an interaction between an agent and the KG environment: starting from the topic entity, the agent incrementally extends a path by selecting relations and entities, and optimizes its policy using a termination signal that indicates whether the answer entity has been reached (Nie et al., 2024; Jiang et al., 2023). A representative work ARN

* Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 22870–22884

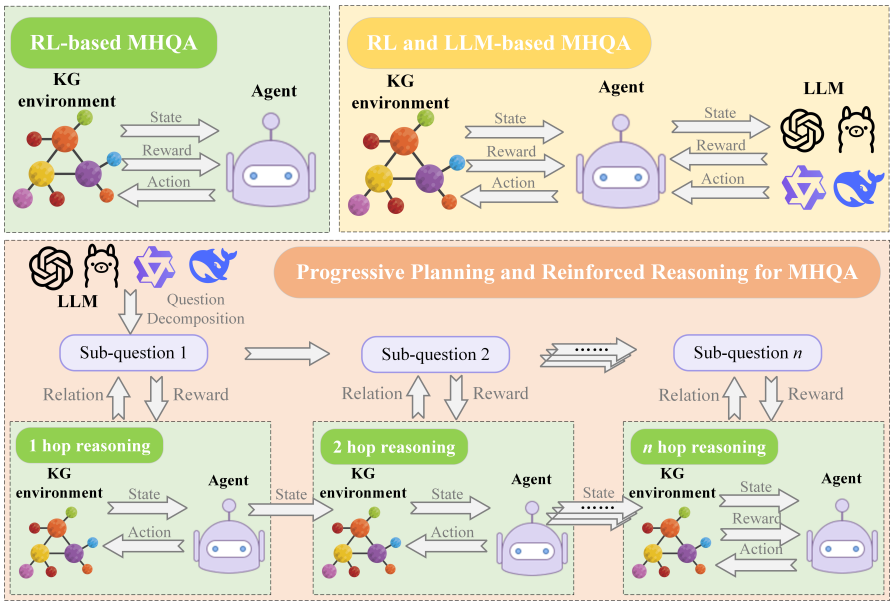

Figure 2: From RL to RL+LLM to the proposed framework. (a) RL-based MHQA: The agent searches for answer paths on the KG via a state–action–reward interaction loop, but is prone to undirected exploration. (b) RL and LLM-based MHQA: The LLM provides priors to shape inter-mediate rewards or evaluate candidate actions, thereby reducing blind search. (c) Progressive Planning and Reinforced Reasoning: The LLM decomposes the original question into an ordered sequence of sub-questions, and through subquestion-guided stepwise decision, enables well-grounded multi-hop reasoning.

(Cui et al., 2023c) incorporates KG embeddings as "expected information" within the RL framework, which mitigates issues like sparse and delayed rewards to some extent. Nevertheless, the inherent limitations of RL remain pronounced: the lack of explicit guidance for intermediate reasoning steps often leads the agent to learn shallow heuristics or spurious correlations (Zhang and Zhao, 2025; Hao et al., 2025a). Figure 1 presents the average Hits@1 results over 15 runs. It is evident that the RL-based MHQA not only converges more slowly and to a lower peak performance, but also exhibits a rise-then-decline pattern in the mid-to-late training stages. We argue that this instability primarily stems from the lack of explicit intermediate supervision. Without guided feedback at critical reasoning steps, the agent can be easily misled by coincidental successes, ultimately failing to develop robust reasoning strategies and leading to performance collapse.

In recent years, hybrid frameworks that couple large language models (LLMs) with RL have offered fresh avenues for multi-hop reasoning (Zhang and Zhao, 2025; Song et al., 2025): they leverage LLM prior knowledge to generate intermediate signals (e.g., reward shaping, action evaluation), thereby mitigating the disadvantage of pure RL's aimless exploration (see Fig. 1). However, these methods still have notable limitations: first, they typically intervene in decision-making through coarse-grained "indirect signals", while lacking step-by-step guidance over concrete action paths, making it difficult for the agent to acquire an executable multi-hop reasoning policy. Second, existing RL approaches seldom learn the multi-hop subgraph structures of KGs, and cannot predict the potential consequences of decisions in advance, causing the agent to drift into inefficient exploration on complex multi-hop queries.

To this end, we propose a Progressive Planning and Reinforced Reasoning (PPRR) framework. Figure 2 contrasts our approach with RL and RL+LLM pipelines: the proposed framework leverages an LLM to decompose the input question into an ordered sequence of sub-questions, and via a "subquestion-guided stepwise decision" scheme, tightly aligns high-level semantic planning with hop-by-hop search over the knowledge graph, thereby maintaining fine-grained, progressive guidance. This design rests on a simple observation: complex real-world questions can naturally be broken into multiple steps, and each subquestion aligns semantically with relation paths in the KG. Additionally, we further introduce a lookahead policy network to enhance the agent's global situational awareness and decision foresight. It inte-

grates an improved graph-attention module that explicitly encodes inter-node dependencies on multi-hop subgraphs and computes lookahead scores for candidate actions, enabling the policy to prioritize moves with long-term value. As the detailed case in Figure 3 shows, at the second hop, although most movies match the local context of this hop, from the perspective of global answer derivation they are preemptively deemed low-reward by the policy network because they lack an effective path (in_language) to the target. Main contributions are summarized as follows:

(1) We propose a Progressive Planning and Reinforced Reasoning framework for MHQA over knowledge graph. By tightly coupling LLM macro planning with RL micro reasoning, the framework guides agents toward human-style, stepwise decision-making and problem solving.

(2) We develop a structure-aware lookahead policy network that encourages learning global reasoning beyond immediate rewards, substantially reducing ineffective exploration.

(3) We conduct extensive experiments on four public MHQA benchmarks and an additional in-domain dataset. The results show that our framework outperforms state-of-the-art baselines while exhibiting strong generalization.

2 Related Work

This section reviews the strengths and limitations of existing MHQA methods from two perspectives: RL-based, RL and LLM-based MHQA.

2.1 RL-based MHQA

Reinforcement learning models multi-hop question answering as a sequential decision-making problem on a knowledge graph (Kaiser et al., 2021; Hou et al., 2021). This approach not only ensures the interpretability of the reasoning process, but also endows the model with robustness under data sparsity, providing a key technical pathway for building transparent and reliable QA systems (Edwards et al., 2022; Han et al., 2025). Early work by Qiu et al. (2020) first applied RL to MHQA and demonstrated its efficacy across multiple datasets. Building on this, ARL (Zhang et al., 2022) and HRN (Zhang et al., 2025b) seek to alleviate sparse and delayed rewards by designing an adaptive path generator and introducing hierarchical rewards, respectively. Moreover, to address the spurious-path issue, Cui et al. (2023b) proposed an adversarial RL formulation for MHQA. Despite these advances in end-to-end path learning, two common shortcomings remain: (1) the absence of strong priors leaves agents prone to unguided exploration in vast search spaces; and (2) the policy network fails to fully consider the global structural information of subgraphs, preventing adequate perception of subgraph context before making decisions and thereby reducing reasoning accuracy.

2.2 RL and LLM-based MHQA

LLMs serve as knowledge engines that inject commonsense and priors into reinforcement learning (Wang et al., 2024a; Rashidi Laleh and Nili Ahmadabadi, 2024; Xu et al., 2024a). Accordingly, an important research paradigm is to use LLMs to provide structured guidance for each step of an agent's decision-making, thereby fundamentally alleviating the intrinsic challenges of blind exploration and reward sparsity (Hao et al., 2025b; Wang et al., 2024b). For example, Zhang and Zhao (2025) propose an LLM–RL collaborative framework that uses LLM outputs as intermediate rewards to ease sparsity and treats the LLM as an action advisor to score candidate actions, thereby reducing blind exploration. Nevertheless, the guidance in existing methods largely remains at an indirect, post-hoc evaluation level, making it difficult to impose fine-grained, feedforward constraints on specific action paths. Moreover, reliance on LLMs can introduce risks such as hallucination, which reduces the overall accuracy of MHQA.

3 Approach

To address the above challenges, we propose Progressive Planning and Reinforced Reasoning (PPRR) framework for MHQA. As illustrated in Figure 3, the framework introduces two key innovations: (1) Progressive planning. We align LLM-decomposed sub-questions with hop-by-hop decisions on the KG, enabling targeted and subquestion-guided reasoning rather than unguided exploration. (2) Structure-aware lookahead policy network. We explicitly model relational information over multi-hop subgraphs and perform lookahead value evaluation for candidate actions, proactively suppressing ineffective exploration before decisions are made. We now detail each component of PPRR.

3.1 Progressive Planning

LLMs are capable of decomposing complex questions into executable inter-mediate steps (Li et al.,

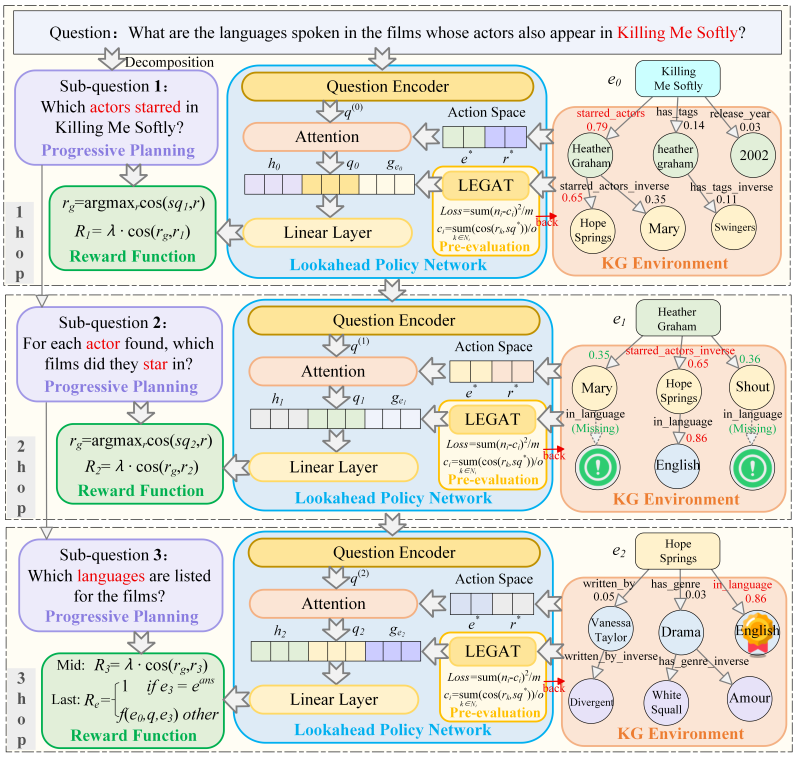

Figure 3: The entire reasoning process of the proposed framework. The LLM decomposes the question into sub-questions. The RL agent navigates the KG step by step, guided by sub-question alignment and the LEGAT lookahead network, which evaluates the long-term utility of actions.

2024b; Chen et al., 2024; Zhang et al., 2024). Given a question q, a topic entity e_0, and the multi-hop sub-graph around e_0, the LLM adaptively produces an ordered sequence of sub-questions $\{sq_1, sq_2, \ldots, sq_n\}$. For example, for the input question "What are the languages spoken in the films whose actors also appear in Killing Me Softly?" the LLM outputs three sub-questions: (1) Which actors starred in Killing Me Softly? (2) For each actor found, which films did they star in? (3) Which languages (in_language) are listed for those films? These sub-questions align one-to-one with the three-hop reasoning from the topic entity Killing Me Softly: Actor → Film → Language. In our implementation, GPT-4 (OpenAI et al., 2024) serves as the planner that generates the sub-question sequence. It should be noted that, in order to cultivate the model's progressive planning ability, sub-question decomposition is used only during the training phase. In addition, we further enforce a KG-alignment validation: whenever any sub-question cannot be matched to a sufficiently similar relation in the knowledge graph (when the similarity falls below the threshold τ= 0.6), we trigger a re-decomposition prompt to mitigate hallucination-induced misguidance from the LLM.

3.2 Reinforced Reasoning

The core of KGQA is to locate answers by exploiting structured knowledge, while RL offers a dynamic framework for path exploration. We cast QA as a Markov decision process (MDP) (BELLMAN, 1957) in which states, actions, and rewards interact to enable stepwise reasoning over the KG and ultimately identify the correct answer. Details are as follows:

State. At time step t, we define the state as $S_t = (q, e_0, e_t, h_t)$, where q denotes the encoded question, e_0 is the topic entity, e_t is the current entity, and h_t is the history of traversed relations, updated via $h_t = \text{LSTM}(h_{t-1}, A_{t-1})$ (Greff et al., 2017). All vectors are initialized to zero at t=0.

Action. The action space comprises outgoing edges from the current entity e_t. Concretely, $A_t = \{(r, e)|(e_t, r, e) \in G\}$, i.e., the agent chooses a relation and entity pair (r, e) from the neighbors of e_t as the next hop.

Reward. The reward combines an intermediate guidance signal and a terminal payoff. For hop h, we first encode the sub-question representation sq_h and select the most compatible guiding rela-

tion r_g from the candidate set. We then compute the cosine similarity between the actually chosen relation r_h and r_g to form the per-hop reward R_h. The detailed calculation process is shown in Eqs.(1-2). This intermediate signal encourages semantic consistency between the selected relation and the current sub-question, thereby suppressing blind exploration during training. All text/relation vectors are produced by a pretrained BERT encoder (Koroteev, 2021).

$$r_g = argmax_r \cos(sq_h, r) \tag{1}$$

$$R_h = \lambda \cdot \cos(r_g, r_h) \tag{2}$$

At the terminal step T, the final reward is computed via Eq.(3). To mitigate sparsity from binary success/failure signals, we utilize the scoring function f() (Cui et al., 2023c) of the KGE model to compute soft rewards for candidate entities. This soft metric provides denser and more stable supervision for policy learning, thereby mitigating reward sparsity and improving training stability.

$$R_e = \begin{cases} 1, & \text{if } e_T = e^{ans} \\ f(e_0, q, e_T), & \text{otherwise} \end{cases} \tag{3}$$

where e_T is the predicted entity and e^{ans} is the golden answer.

3.3 Structure-Aware Lookahead Policy Network

As the core execution component of MHQA, the policy network maps the current state to a probability distribution over candidate actions at each hop. Specifically, we encode the input question q using a Transformer encoder (Vaswani et al., 2017). Considering that different time steps t may focus on different parts of the question, we employ a linear layer to generate stepwise vector representations $q^{(t)}$. The implementation details are provided in Eq.(4).

$$q^{(t)} = [q_1^{(t)}, q_2^{(t)}, \ldots, q_n^{(t)}] = \text{Tanh}(W_t \cdot q + b_t) \tag{4}$$

where q_i is the vector representation of the i-th token of q, Tanh is the activation function, and W_t and b_t are learnable parameter matrices. Next, given the state $S_t = (q, e_0, e_t, h_t)$ and the action space A_t, we compute the relation-aware question representation q_t using Eqs.(5-6), thereby learning token-level alignments between the relation and the question for each action $A^* = (r^*, e^*) \in A_t$.

$$q_t = \sum_{i=1}^{n} \alpha_i \cdot q_i^{(t)} \tag{5}$$

$$\alpha_i = \frac{\exp(W_a \cdot (r^* \bigodot q_i^{(t)}))}{\sum_{j=1}^{n} \exp(W_a \cdot (r^* \bigodot q_j^{(t)}))} \tag{6}$$

where W_a is a learnable parameter matrix, and r^* represents the relation vector generated by the pretrained ConvE model (Dettmers et al., 2018). To endow the policy network with deeper insight, we propose Graph Attention Networks with Lookahead Evaluation (LEGAT). Traditional graph attention mechanisms lack a long-term perspective when assessing the importance of connections, which can cause agents to become trapped in locally optimal exploration. LEGAT overcomes this limitation by explicitly encoding lookahead evaluation into the message-passing process. Concretely, at layer l, LEGAT updates each node representation g_j according to Eq.(7).

$$g_j^{l+1} = f_m(\sum_{i \in N_j} n_i m_{ij}) + g_j^{(l)} \tag{7}$$

where N_j denotes the neighborhood of node j. The message m_{ij} is passed from node i to node j and consists of the node embedding g and the relation embedding r (i.e., the action space). Note that these embeddings are produced by a pretrained ConvE model (Dettmers et al., 2018). The function f_m aggregates incoming messages via a two-layer MLP and normalization (Ioffe and Szegedy, 2015). Unlike standard graph attention (Veličković et al., 2018), we replace raw attention weights with a lookahead score n_i for each action/message, computed as in Eq.(8). This score, computed via a lightweight neural network, aims to capture the multi-step value of message transmission from node i to node j, rather than solely relying on the immediate relevance of their current states.

$$n_i = \underset{i \in N_j}{\text{Tanh}}(W_b \cdot [g_i, m_{ij}] + b_b) \tag{8}$$

where W_b and b_b are learnable parameters and Tanh denotes the activation function. To cultivate the network's lookahead evaluation capability, a lookahead evaluation loss is proposed and jointly optimized with the reasoning loss. It supervises the lookahead score n to fit a lookahead value c, as detailed in Eqs.(9-10). The computation of c is strategically designed: it evaluates not only the relevance between the candidate relation and the associated sub-question, but also the anticipated quality of the outgoing (next-hop) relations from the candidate node. This design enables the model to identify and

suppress paths that appear locally relevant but cannot reach the target, thereby achieving pre-decision pruning and reducing blind exploration. It is worth noting that although the lookahead value c is defined only over the one-hop neighbors of a candidate node, LEGAT's message-passing mechanism can progressively propagate and accumulate the expectation signal. Consequently, higher-layer representations can integrate lookahead information from more distant regions of the subgraph, thereby enabling structural multi-hop awareness and global evaluation of potential reasoning paths.

$$Loss = \frac{1}{m} \cdot \sum_{i=1}^{m} (n_i - c_i)^2 \tag{9}$$

$$c_i = \frac{1}{o} \cdot \sum_{k \in N_i} \cos(r_k, sq^*) \tag{10}$$

As before, we encode the relations and sub-questions using a pretrained BERT encoder (Koroteev, 2021). Here, N_i denotes the neighborhood of the candidate node i. Additionally, m denotes the number of nodes in the local subgraph, and o denotes the number of relations adjacent to node i. We use sq^* to indicate the sub-question aligned to relation r_k. Concretely, if the hop distance from the current entity e_t to r_k is n, then sq^* refers to the $(t+n)$-th sub-question. If $t+n$ exceeds the number of sub-questions, those relations are excluded from the computation of c.

As illustrated in Fig.3, the lookahead score becomes pivotal at the second hop, preemptively flagging film candidates that satisfy actor co-appearance yet lack the in_language relation required by the final objective. This exemplifies the proposed framework's distinctive foresight in complex multi-hop reasoning: it proactively prunes trajectories that cannot reach the target. Finally, the policy network selects the most promising action based on the path history ht, the relation-aware question representation q_t, and lookahead structural cue g_{et}, as detailed in Eq.(11).

$$\pi(a_t|S_t) = A^* \cdot W_c \cdot \text{ReLU}(W_d \cdot [h_t; q_t; g_{e_t}]) \tag{11}$$

where W_c and W_d are learnable parameter matrices, and ReLU serves as the activation function. Although both LEGAT and ARN (Cui et al., 2023c) introduce future-oriented signals to mitigate blind exploration in RL-based KGQA, their mechanisms differ fundamentally. ARN is answer-oriented: it derives an anticipation embedding from KGE-based scoring over candidate answer entities and injects this embedding into the RL state as a global target prior. In contrast, LEGAT is action-centric and structure-aware: it assigns a lookahead score to each candidate action through graph message passing, without explicitly inferring a target-entity embedding. Moreover, LEGAT's lookahead scores are trained with an explicit supervision objective that captures both current sub-question alignment and the expected quality of subsequent transitions. Therefore, ARN provides an answer-level global prior, whereas LEGAT performs a structured, fine-grained evaluation of candidate actions during reasoning.

3.4 Optimization and Training

During training, we follow Zhang and Zhao (2025) and optimize the policy network using the REINFORCE algorithm. Let the training set be D. Our objective is to maximize the expected cumulative return over question–answer pairs (q, a), as formalized in Eq.(12). The gradient of the objective $J(\theta)$ with respect to the policy parameters θ, given in Eq.(13), guides the policy updates to maximize the expected return, thus progressively improving the policy network.

$$J(\theta) = E_{(q,a)\in D}[E_{a_1,a_2,\ldots,a_T\sim\pi}[\sum_{t=1}^{T} R(S_t)]] \tag{12}$$

$$\nabla J(\theta) = E_{(q,a)\in D}[E_{a_1,a_2,\ldots,a_T\sim\pi}[\sum_{t=1}^{T} R(S_t)\nabla \log \pi]] \tag{13}$$

4 Experiment

4.1 Datasets and Evaluation Metrics

To assess the effectiveness of our framework, we evaluate it on four public MHQA benchmarks: WebQuestionSP (WebQSP) (Yih et al., 2016), PathQuestion (PQ) (Zhou et al., 2018), PathQuestion-Large (PQL) (Zhou et al., 2018) and MoviE Text Audio QA (MetaQA) (Zhang et al., 2018). These datasets are also among the most widely used in existing RL-based MHQA research. For each dataset, we adopt both the most challenging subsets (i.e., those with the largest hop) and mixed-hop subsets to rigorously probe complex reasoning ability. Performance is reported using the standard Hits@1, which measures whether the top-ranked prediction matches the gold answer. We give a detailed description of each dataset in Appendix A.

Methods	WebQSP	PQ		PQL		MetaQA
	Mix	Mix	3H	Mix	3H	3H
KVMemNN (Miller et al., 2016)	46.7	79.4	85.2	63.4	68.6	53.8
IRN (Zhou et al., 2018)	–	83.3	85.8	61.8	62.4	35.6
EmbedKGQA (Saxena et al., 2020)	66.6	–	–	–	–	94.8
DRN (Cui et al., 2023a)	–	92.3	93.5	95.2	92.8	95.9
BSEM(Yuan et al., 2024)	70.8	–	–	–	–	–
PullNet(Sun et al., 2019)	68.1	–	–	–	–	91.4
2HR-DR(Han et al., 2020)	67.0	–	–	92.1	–	–
HyperTransformer(Heo et al., 2022)	–	90.3	89.5	95.4	94.5	–
ComPath(Li et al., 2024a)	–	91.5	–	–	–	84.4
SRN(Qiu et al., 2020)	–	89.2	89.3	77.5	78.3	75.2
ARN(Cui et al., 2023c)	68.0	90.6	93.7	94.2	95.0	97.0
AR2N(Cui et al., 2023b)	–	92.3	93.9	95.2	95.8	96.2
HRN(Zhang et al., 2025b)	–	93.1	95.3	–	–	–
SCR(Han et al., 2025)	–	94.0	94.0	97.1	98.5	95.0
CRF(Zhang and Zhao, 2025)	79.5	95.7	97.1	–	–	99.2
PPRR (Ours)	**84.5**	**96.9**	**98.5**	**98.1**	**99.2**	**99.5**

Table 1: Performances (%Hits@1) of different MHQA methods. The best score is in **bold**, the second-best is underlined, and "–" indicates no results were reported in the original papers.

4.2 Baselines

To provide a comprehensive evaluation, we group baselines into three families: (1) Neural network-based methods: KVMemNN (Miller et al., 2016), IRN (Zhou et al., 2018), EmbedKGQA (Saxena et al., 2020), DRN (Cui et al., 2023a), BSEM (Yuan et al., 2024). (2) Graph neural network-based methods: PullNet (Sun et al., 2019), 2HR-DR (Han et al., 2020), HyperTransformer (Heo et al., 2022), Compath (Li et al., 2024a). (3) Reinforcement learning-based methods: SRN (Qiu et al., 2020), ARN (Cui et al., 2023c), AR2N (Cui et al., 2023b), HRN (Zhang et al., 2025b), SCR (Han et al., 2025), CRF (Zhang and Zhao, 2025). The detailed description is introduced in Appendix C.

4.3 Main Results

Table 1 summarizes results on four public MHQA benchmarks: WebQSP (Yih et al., 2016), PQ (Zhou et al., 2018), PQL (Zhou et al., 2018), and MetaQA (Zhang et al., 2018). Overall, our framework achieves the best performance on all datasets, substantiating the effectiveness of PPRR. In particular, neural network and graph neural network-based methods generally lag behind RL-based approaches, indicating that the interactive nature of RL is better suited to path search and reasoning over knowledge graphs. In contrast, pure neural or graph models lack explicit stepwise exploration and struggle to capture the path dependencies in multi-hop reasoning. That said, conventional RL methods often suffer from aimless exploration and low-quality reward signals, which limit their performance.

Marrying LLM priors with RL effectively pushes past this ceiling. Notably, recent approaches such as CRF (Zhang and Zhao, 2025) achieve substantial gains across datasets, highlighting the synergistic advantage of combining RL's path exploration capability with the semantic priors of LLMs. Building on this line, our framework delivers a further 5% improvement on the most challenging WebQSP (Yih et al., 2016) benchmark, affirming its suitability and superiority for KG-based MHQA. This advance stems from two design pillars: (1) progressive planning, which tightly fuses the LLM's macro-level decomposition with the RL agent's micro-level path search, thereby strengthening directional guidance throughout multi-hop reasoning; and (2) a structure-aware lookahead policy network that precisely identifies critical nodes and viable connections along the reasoning path, proactively avoiding off-target decisions and further improving path-search accuracy.

4.4 Comparison with LLM-based Methods

As shown in Table 2, PPRR demonstrates a favorable accuracy–efficiency trade-off when compared with competitive LLM-based baselines. On WebQSP, its Hits@1 score (84.5) is lower than that

Methods	WebQSP	MetaQA-3H	Times(s)
ToG w/GPT-4 (Sun et al., 2024)	82.6	-	63.1
PoG w/GPT-4 (Chen et al., 2024)	87.3	-	16.8
ReKnoS w/GPT-4o-mini (Wang et al., 2025b)	83.8	-	3.7
DoG w/GPT-4 (Ma et al., 2025)	91.0	96.0	8.5
PPRR w/GPT-4	84.5	99.5	0.17

Table 2: Results of different LLM-based methods.

of the strongest prompting-agent methods, such as DoG (91.0) and PoG (87.3), but its inference time is only 0.17 seconds, which is substantially faster than all LLM-based competitors. On MetaQA-3H, PPRR achieves the best performance (99.5), outperforming DoG (96.0) while maintaining the same low latency. This comparison highlights that PPRR internalizes the planning capability of LLMs into a reasoning model, thereby eliminating the need for repeated LLM calls at inference time and achieving a more practical balance between accuracy and efficiency.

4.5 Ablation Study

Methods	WebQSP	PQ	PQL	MetaQA	Average Steps
	Mix	Mix	Mix	3H	
PPRR	84.5	98.5	99.2	99.5	1.93
w/o PP	77.9	94.9	95.5	96.4	2.16
w/o LEGAT	81.1	96.8	97.0	97.7	2.09

Table 3: Ablation experiments of proposed framework.

Table 3 reports the ablation results of the PPRR framework. By contrasting the full model with variants that remove key components, we directly verify the necessity and contribution of both Progressive Planning(PP) and the LEGAT-based policy network. It should be noted that the average steps refer to the mean number of reasoning steps per test sample across all four datasets. Specifically, removing progressive planning leads to substantial degradation across all datasets. On the most challenging WebQSP (Yih et al., 2016) benchmark, accuracy declines by 6.6%. This indicates that without subquestion-guided direction, the model reverts to the unguided random exploration of vanilla RL, severely undermining reasoning accuracy and efficiency. In addition, removing LEGAT also yields noticeable declines in QA accuracy and increases in reasoning steps, confirming that LEGAT explicitly models multi-hop subgraph structures and anticipates the long-term value of actions, thereby enhancing the RL model's global awareness and decision-making foresight.

4.6 Efficiency Evaluation

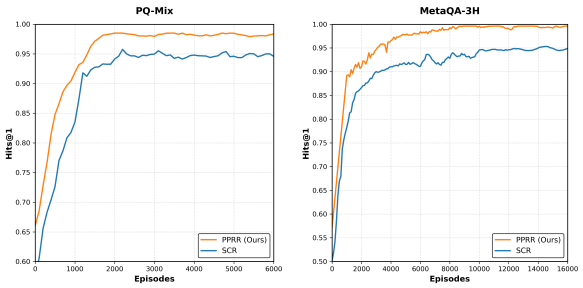

Figure 4: Convergence curves on different datasets.

To further assess convergence efficiency, we compare learning curves on PQ-Mix (Zhou et al., 2018) and MetaQA-3H (Zhang et al., 2018), as shown in Figure 4. Across both datasets, PPRR consistently surpasses the baseline SCR (Han et al., 2025) in terms of initial accuracy, convergence speed, and ultimate performance. These results provide strong evidence that, by deftly coupling the strengths of LLMs and RL, our framework achieves a more efficient training process and superior outcomes for multi-hop reasoning over knowledge graphs.

4.7 Case Study

As shown in Figure 5, the case study clearly illustrates the effectiveness of our framework. It should be noted that this case is drawn from the test set, meaning the evaluation was conducted without guidance from a large language model. The results show that the framework's reasoning path remains precisely aligned with the sequence of sub-questions, confirming that during training the agent successfully internalized the "progressive planning" mindset as its own reasoning capability.

More importantly, the lookahead policy network exhibits decision foresight. At the first hop, the agent faces multiple candidate actions. Although Up and Paris are consistent with the current subquestion, the prospective network early identifies their lack of the necessary release relationship for the next step by evaluating their long-term value, thereby assigning them lower weights. By contrast, it assigns the highest weight to Venus, whose path leads successfully to the correct answer Deodato_2. This demonstrates that the lookahead mechanism can preemptively filter out paths that are "locally correct but globally ineffective", enabling pre-execution pruning that significantly improves both efficiency and success rate in multi-hop

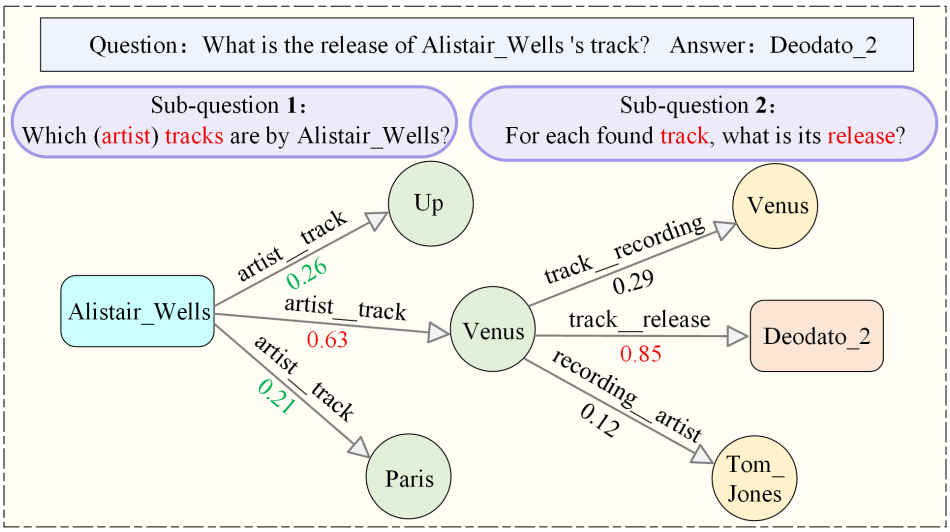

Figure 5: A case from PQL-2H dataset.

reasoning.

4.8 Domain Applicability Evaluation

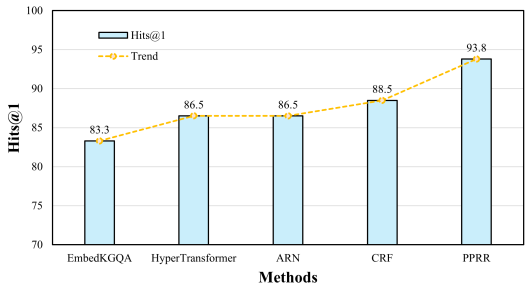

Figure 6: Performances on Power Domain Dataset.

To further assess the framework's generalization ability, we conduct experiments on a power-domain QA dataset. It should be noted that this dataset also includes reasoning tasks with different hop counts on the domain knowledge graph. The most challenging of these is the fault diagnosis question-answering task with more than 3 hops, which involves multiple types of nodes such as fault phenomena, fault types, fault causes, and fault locations. Specifically, we randomly split 962 fault-related QA pairs into training, validation and test sets with an 8:1:1 ratio and evaluate using Hits@1 metric. Figure 6 illustrates the results of representative baselines on this in-domain benchmark. Our framework consistently outperforms competing methods, indicating strong domain adaptability. We attribute these gains to the structure-aware lookahead policy network, which accurately captures fault-related dependencies in the power KG and thereby strengthens reasoning over complex fault-propagation paths.

5 Conclusion

To address issues such as blind exploration in existing reinforcement learning methods for MHQA on knowledge graphs, we propose a Progressive Planning and Reinforced Reasoning framework. Subquestion-guided stepwise decisions give the agent human-like stepwise problem-solving cognition, and a structure-aware lookahead policy network estimates the long-term return of multi-hop reasoning paths. Experiments on multiple public benchmarks and a domain-specific dataset demonstrate that PPRR not only delivers substantial performance gains but also curbs ineffective exploration and backtracking via proactive decision-making. These results validate the framework's effectiveness and generalization in complex multi-hop reasoning. In future work, we will further enhance generalization and robustness to better handle sparse KGs and stringent logical constraints.

Limitations

Although the proposed framework achieves strong performance on multiple MHQA datasets, its application scope can be further broadened. First, the current planning and reasoning efficacy still depends to some extent on the structural completeness of the underlying KG. Second, tackling complex logic-intensive queries (e.g., multi-constraint comparative questions) calls for a finer-grained coordination mechanism between semantic planning and action execution. Third, the framework's generalization to unseen relations has not been sys-

tematically evaluated in this work and remains an important direction for future research.

Acknowledgements

We are grateful to the anonymous reviewers and the area chair for their insightful comments and suggestions. This work was supported in part by BUPT Kunpeng and Ascend Center of Cultivation. This work was also supported by the National Natural Science Foundation of China (Grant No. 62373150).

References

RICHARD BELLMAN. 1957. A markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684.

Sheng Bi, Zeyi Miao, and Qizhi Min. 2025. A modular dual learning for improving question answering and generation over knowledge graphs. *IEEE Transactions on Audio, Speech and Language Processing*, 33:401–417.

Jiabei Chen, Guang Liu, Shizhu He, Kun Luo, Yao Xu, Jun Zhao, and Kang Liu. 2025. Search-in-context: Efficient multi-hop QA over long contexts via Monte Carlo tree search with dynamic KV retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26443–26455, Vienna, Austria. Association for Computational Linguistics.

Liyi Chen, Panrong Tong, Zhongming Jin, Ying Sun, Jieping Ye, and Hui Xiong. 2024. Plan-on-graph: Self-correcting adaptive planning of large language model on knowledge graphs. In *Advances in Neural Information Processing Systems*, volume 37, pages 37665–37691. Curran Associates, Inc.

Hai Cui, Tao Peng, Tie Bao, Ridong Han, Jiayu Han, and Lu Liu. 2023a. Stepwise relation prediction with dynamic reasoning network for multi-hop knowledge graph question answering. *Applied Intelligence*, 53(10):12340–12354.

Hai Cui, Tao Peng, Ridong Han, Jiayu Han, and Lu Liu. 2023b. Path-based multi-hop reasoning over knowledge graph for answering questions via adversarial reinforcement learning. *Knowledge-Based Systems*, 276:110760.

Hai Cui, Tao Peng, Feng Xiao, Jiayu Han, Ridong Han, and Lu Liu. 2023c. Incorporating anticipation embedding into reinforcement learning framework for multi-hop knowledge graph question answering. *Information Sciences*, 619:745–761.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Gavin Edwards, Sebastian Nilsson, Benedek Rozemberczki, and Eliseo Papa. 2022. Explainable biomedical recommendations via reinforcement learning reasoning on knowledge graphs. *Preprint*, arXiv:2111.10625.

Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. 2017. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28:2222–2232.

Jiale Han, Bo Cheng, and Xu Wang. 2020. Two-phase hypergraph based reasoning with dynamic relations for multi-hop kbqa. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3615–3621. International Joint Conferences on Artificial Intelligence Organization. Main track.

Ridong Han, Jia Liu, Haijia Bi, Tao Peng, and Lu Liu. 2025. Scr: A completion-then-reasoning framework for multi-hop question answering over incomplete knowledge graph. *Neurocomputing*, 651:131027.

Chuzhan Hao, Wenfeng Feng, Yuewei Zhang, and Hao Wang. 2025a. Dynasearcher: Dynamic knowledge graph augmented search agent via multi-reward reinforcement learning. *Preprint*, arXiv:2507.17365.

Qianyue Hao, Yiwen Song, Qingmin Liao, Jian Yuan, and Yong Li. 2025b. Llm-explorer: A plug-in reinforcement learning policy exploration enhancement driven by large language models. *Preprint*, arXiv:2505.15293.

Yu-Jung Heo, Eun-Sol Kim, Woo Suk Choi, and Byoung-Tak Zhang. 2022. Hypergraph transformer: Weakly-supervised multi-hop reasoning for knowledge-based visual question answering. *Preprint*, arXiv:2204.10448.

Zhongni Hou, Xiaolong Jin, Zixuan Li, and Long Bai. 2021. Rule-aware reinforcement learning for knowledge graph reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4687–4692, Online. Association for Computational Linguistics.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.

Yixin Ji, Kaixin Wu, Juntao Li, Wei Chen, Mingjie Zhong, Xu Jia, and Min Zhang. 2024. Retrieval and reasoning on KGs: Integrate knowledge graphs into large language models for complex question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7598–7610, Miami, Florida, USA. Association for Computational Linguistics.

Chunyang Jiang, Tianchen Zhu, Haoyi Zhou, Chang Liu, Ting Deng, Chunming Hu, and Jianxin Li. 2023.

Path spuriousness-aware reinforcement learning for multi-hop knowledge graph reasoning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3181–3192, Dubrovnik, Croatia. Association for Computational Linguistics.

Yangqin Jiang, Yuhao Yang, Lianghao Xia, and Chao Huang. 2024. Diffkg: Knowledge graph diffusion model for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 313–321, New York, NY, USA. Association for Computing Machinery.

Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement learning from reformulations in conversational question answering over knowledge graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 459–469. ACM.

M. V. Koroteev. 2021. BERT: A review of applications in natural language processing and understanding. *CoRR*, abs/2103.11943.

Sirui Li, Kok Wai Wong, Dengya Zhu, and Chun Che Fung. 2024a. Enhancing question answering through effective candidate answer selection and mitigation of incomplete knowledge graphs and over-smoothing in graph convolutional networks. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. *ACM Trans. Inf. Syst.*, 43(3).

Yading Li, Dandan Song, Changzhi Zhou, Yuhang Tian, Hao Wang, Ziyi Yang, and Shuhao Zhang. 2024b. A framework of knowledge graph-enhanced large language model based on question decomposition and atomic retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11472–11485, Miami, Florida, USA. Association for Computational Linguistics.

Qichuan Liu, Chentao Zhang, Chenfeng Zheng, Guosheng Hu, Xiaodong Li, and Zhihong Zhang. 2025. Beyond the answer: Advancing multi-hop QA with fine-grained graph reasoning and evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23433–23456, Vienna, Austria. Association for Computational Linguistics.

Jie Ma, Zhitao Gao, Qi Chai, Wangchun Sun, Pinghui Wang, Hongbin Pei, Jing Tao, Lingyun Song, Jun Liu, Chen Zhang, and Lizhen Cui. 2025. Debate on graph: A flexible and reliable reasoning framework for large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24768–24776.

Anastasia Martynova, Vladislav Tishin, and Natalia Semenova. 2025. Learn together: Joint multitask finetuning of pretrained KG-enhanced LLM for downstream tasks. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, pages 13–19, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. *Preprint*, arXiv:1606.03126.

Weizhi Nie, Xin Wen, Jing Liu, Jiawei Chen, Jiancan Wu, Guoqing Jin, Jing Lu, and An-An Liu. 2024. Knowledge-enhanced causal reinforcement learning model for interactive recommendation. *IEEE Transactions on Multimedia*, 26:1129–1142.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Piyushkumar Patel. 2025. Graph-enhanced retrieval-augmented question answering for e-commerce customer support. *Preprint*, arXiv:2509.14267.

Yunqi Qiu, Yuanzhuo Wang, Xiaolong Jin, and Kun Zhang. 2020. Stepwise reasoning for multi-relation question answering over knowledge graph with weak supervision. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 474–482, New York, NY, USA. Association for Computing Machinery.

Alireza Rashidi Laleh and Majid Nili Ahmadabadi. 2024. A Survey On Enhancing Reinforcement Learning in Complex Environments: Insights from Human and LLM Feedback. *arXiv e-prints*, arXiv:2411.13410.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. 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*, pages 4498–4507, Online. Association for Computational Linguistics.

Tiesunlong Shen, Jin Wang, Xuejie Zhang, and Erik Cambria. 2025. Reasoning with trees: Faithful question answering over knowledge graph. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3138–3157, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiaozhuang Song, Shufei Zhang, and Tianshu Yu. 2025. ReKG-MCTS: Reinforcing LLM reasoning on knowledge graphs via training-free Monte Carlo tree search. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9288–9306,

Vienna, Austria. Association for Computational Linguistics.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. 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)*, pages 2380–2390, Hong Kong, China. Association for Computational Linguistics.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *Preprint*, arXiv:2307.07697.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. *Preprint*, arXiv:1710.10903.

Boyuan Wang, Yun Qu, Yuhang Jiang, Jianzhun Shao, Chang Liu, Wenming Yang, and Xiangyang Ji. 2024a. Llm-empowered state representation for reinforcement learning. *Preprint*, arXiv:2407.13237.

Jie Wang, Alexandros Karatzoglou, Ioannis Arapakis, and Joemon M. Jose. 2024b. Reinforcement learning-based recommender systems with large language models for state reward and action modeling. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 375–385, New York, NY, USA. Association for Computing Machinery.

Nan Wang, Yongqi Fan, yansha zhu, ZongYu Wang, Xuezhi Cao, Xinyan He, Haiyun Jiang, Tong Ruan, and Jingping Liu. 2025a. Kg-o1: Enhancing multi-hop question answering in large language models via knowledge graph integration. *Preprint*, arXiv:2508.15790.

Song Wang, Junhong Lin, Xiaojie Guo, Julian Shun, Jundong Li, and Yada Zhu. 2025b. Reasoning of large language models over knowledge graphs with super-relations. *Preprint*, arXiv:2503.22166.

Yihong Wu, Liheng Ma, Muzhi Li, Jiaming Zhou, Lei Ding, Jianye Hao, Ho fung Leung, Irwin King, Yingxue Zhang, and Jian-Yun Nie. 2025. Advancing multi-agent rag systems with minimalist reinforcement learning. *Preprint*, arXiv:2505.17086.

Hao Xu, Yunxiao Zhao, Jiayang Zhang, Zhiqiang Wang, and Ru Li. 2025. LOG: A local-to-global optimization approach for retrieval-based explainable multi-hop question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9085–9095, Abu Dhabi, UAE. Association for Computational Linguistics.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024a. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 1362–1373, New York, NY, USA. Association for Computing Machinery.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024b. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2905–2909, New York, NY, USA. Association for Computing Machinery.

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. 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)*, pages 201–206.

Mingcai Yuan, Qiang Lu, Xianhao Zeng, Jake Luo, and Dawei Li. 2024. Alleviating semantic drift in multi-hop question answering on knowledge graphs with bidirectional semantics. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Ding-Chu Zhang, Yida Zhao, Jialong Wu, Liwen Zhang, Baixuan Li, Wenbiao Yin, Yong Jiang, Yu-Feng Li, Kewei Tu, Pengjun Xie, and Fei Huang. 2025a. EvolveSearch: An iterative self-evolving search agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13134–13147, Suzhou, China. Association for Computational Linguistics.

Kun Zhang, Jiali Zeng, Fandong Meng, Yuanzhuo Wang, Shiqi Sun, Long Bai, Huawei Shen, and Jie Zhou. 2024. Tree-of-reasoning question decomposition for complex question answering with large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Qixuan Zhang, Xinyi Weng, Guangyou Zhou, Yi Zhang, and Jimmy Xiangji Huang. 2022. Arl: An adaptive reinforcement learning framework for complex question answering over knowledge base. *Information Processing & Management*, 59(3):102933.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Zhiqiang Zhang, Zhiyi Zhang, Yunxiao Zhang, and Wen Zhao. 2025b. A hierarchical reasoning framework for complex question answering over knowledge graph with reinforcement learning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zhiqiang Zhang and Wen Zhao. 2025. A collaborative reasoning framework powered by reinforcement learning and large language models for complex questions answering over knowledge graph. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10672–10684, Abu Dhabi, UAE. Association for Computational Linguistics.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2018. An interpretable reasoning network for multi-relation question answering. *Preprint*, arXiv:1801.04726.

Yujia Zhou, Zheng Liu, and Zhicheng Dou. 2025. How credible is an answer from retrieval-augmented LLMs? investigation and evaluation with multi-hop QA. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4232–4242, Abu Dhabi, UAE. Association for Computational Linguistics.

A Datasets

Datasets		KG	Train	Valid	Test
WebQSP	Mix	Freebase	2848	260	1639
PQ	3H	Freebase	4163	535	520
	Mix	Freebase	5691	704	711
PQL	3H	Freebase	825	102	104
	Mix	Freebase	2101	260	264
MetaQA	3H	OMDb	114196	14274	14274

Table 4: Statistics of the experiment datasets.

Table 4 summarizes the statistics of the datasets used in our experiments, listing the dataset names, associated knowledge graphs, question types, and the numbers of training/validation/test instances. Most datasets are built on the Freebase knowledge graph; MetaQA uses OMDb and has a training set substantially larger than the others.

B Implementation Details

In all experiments, we use 300-dimensional pre-trained GloVe embeddings for tokens, and set both entity and relation embedding sizes to 200. The LSTM (Greff et al., 2017) encoder for decision history uses a 200-dimensional hidden state. For each pre-trained Transformer (Vaswani et al., 2017) question encoder, we employ 2 layers with 4 attention heads. Moreover, the batch size is 32 and the learning rate is 0.0001. The intermediate-reward coefficient λ is selected from 0.2,0.5,0.7,1.0. In addition, we incorporate a λ decay strategy, with detailed experimental analysis provided in Appendix D. The number of graph-network layers l is set equal to the maximum hop count of questions in the dataset. All experiments are performed on a Linux-based server with 2 NVIDIA A40 GPUs.

C Baselines

KVMemNN (Miller et al., 2016) is a key–value memory neural network that uses the key for addressing and the value for reading, and outputs an answer after accumulating evidence through multi-hop reads.

IRN (Zhou et al., 2018) incrementally parses the question via multi-hop reasoning and predicts relations in the knowledge base, thereby constructing a traceable path to the an-swer.

EmbedKGQA (Saxena et al., 2020) directly predicts the answer entity by mapping the question and the entities in the knowledge graph into a shared vector space.

DRN (Cui et al., 2023a) is a dynamic reasoning network that performs multi-hop knowledge graph question answering by progressively predicting relation paths.

BSEM (Yuan et al., 2024) alleviates the semantic drift caused by path expansion in multi-hop knowledge graph QA by introducing reverse semantic reasoning and incorporating joint and contrastive learning.

PullNet (Sun et al., 2019) constructs a question-relevant heterogeneous subgraph from the knowledge base and text via iterative retrieval, and performs answer inference using a graph convolutional network.

2HR-DR (Han et al., 2020) is a multi-hop knowledge base question answering model built on a directed hypergraph convolutional network. It explicitly updates relation representations and predicts relation paths hop by hop through a two-stage dynamic relation reasoning mechanism, enabling interpretable QA reasoning.

HyperTransformer (Heo et al., 2022) builds hypergraphs for the question and the knowledge, and uses guided attention and self-attention to learn higher-order hypergraph associations, thereby en-

abling knowledge reasoning and QA under weak supervision.

Compath (Li et al., 2024a) is a two-stage framework for multi-hop knowledge graph question answering. It uses CompGCN for knowledge graph completion to mitigate incompleteness, and employs a path analyzer that fuses semantic and structural information to select answers.

SRN (Qiu et al., 2020) models multi-relational knowledge graph question answering as a sequential decision problem, progressively inferring the answer via path search and attention mechanisms.

ARN (Cui et al., 2023c) is a model that integrates expectation embeddings into a reinforcement learning framework, aiming to address issues such as blind exploration and sparse rewards.

AR2N (Cui et al., 2023b) adopts an adversarial learning approach to mitigate the spurious-path problem in RL-based MHQA.

HRN (Zhang et al., 2025b) is a framework based on hierarchical reinforcement learning, which decomposes complex question answering tasks into two processes: high-level constraint detection and low-level path reasoning.

SCR (Han et al., 2025) adopts the idea of "completion then answering" to address the incompleteness of knowledge graphs, and designs semantic rewards to mitigate the sparse reward problem.

CRF (Zhang and Zhao, 2025) is a collaborative reasoning framework that integrates hierarchical reinforcement learning with LLMs, effectively addresses issues such as low exploration efficiency in complex knowledge-graph question answering.

D Hyperparameter Analysis

λ	WebQSP	PQ	PQL	MetaQA
	Mix	Mix	Mix	3H
0.2	82.3	96.8	97.7	98.2
0.5	83.2	97.6	98.5	98.8
0.7	83.5	98.5	98.5	99.2
1	82.8	98.5	99.2	99.5

Table 5: Experiments under different intermediate-reward coefficient.

Table 5 reports the results under different intermediate-reward coefficients λ. Except for WebQSP (Yih et al., 2016), increasing λ from 0.2 to 1.0 yields steady accuracy gains on PQ (Zhou et al., 2018) and PQL (Zhou et al., 2018), MetaQA (Zhang et al., 2018) indicating that strengthening intermediate reward can more effectively guide multi-hop reasoning. In contrast, WebQSP (Yih et al., 2016) peaks at λ=0.7; pushing λ to 1.0 leads to a slight decline. We argue that on relatively difficult datasets, overly strong intermediate rewards can constrain the agent's ability to develop independent and globally-oriented decision-making capabilities during later training stages. Conversely, once the agent has established fundamental reasoning skills through early-stage guidance, "letting go" at the right time can better stimulate its potential for autonomous exploration. Therefore, we adopt a decaying intermediate-reward coefficient strategy. Specifically, during the early stages of training, a relatively large coefficient is used to help the model quickly develop sensitivity to intermediate reasoning targets, guiding it toward effective exploration. As training progresses, the coefficient is gradually reduced, shifting the agent's focus from local subgoals to the global optimization objective and preventing over-reliance on intermediate rewards, as shown in Eq.(14).

$$\lambda = \frac{1}{1 + 0.0003 \times eposides} \tag{14}$$

Under this strategy, the model achieves an optimal accuracy of 84.5%, demonstrating that the decaying mechanism successfully balances short-term exploration and long-term optimization, effectively resolving the performance degradation caused by conflicting reward signals in complex datasets.

E Different LLM Analysis

LLMs	WebQSP	PQ	PQL	MetaQA
	Mix	Mix	Mix	3H
PPRR w/Llama3-8b	81.6	97.3	97.7	98.9
PPRR w/Deepseek	83.8	98.5	98.5	99.4
PPRR w/ChatGPT	82.9	97.5	98.5	99.2
PPRR w/GPT-4	84.5	98.5	99.2	99.5

Table 6: Experiments under different LLMs.

Table 6 presents the experimental results under different LLMs. We observe that even with LLMs of varying capability, the proposed framework maintains strong performance. This further demonstrates that tightly coupling the LLM's reasoning logic with the multi-hop reasoning paths is effective and reduces dependence on any single LLM. Moreover, different LLMs show varying results on the more challenging WebQSP (Yih et al.,

2016) dataset, because GPT-4 can generate more precise sub-questions, thereby mitigating semantic bias or hallucinations.

F Few-shot Study

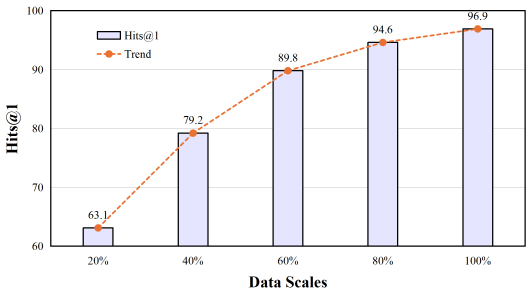

Figure 7: Performances at Different Scales of PQ-3H dataset.

To evaluate few-shot generalization under data scarcity, we vary the fraction of PQ-3H (Zhou et al., 2018) training dataset (20%, 40%, 60%, 80%, 100%) and present results in Figure 7. Even with only 20% of the training set, PPRR attains 63.1% Hits@1. Performance improves steadily as more data are provided, and with 60% of the data the model already approaches the saturated performance of several baselines. These findings substantiate the core advantage of progressive planning: integrating large language models' commonsense priors and planning capabilities into the proposed framework markedly improves learning efficiency and sample utilization, thereby substantially reducing reliance on large labeled datasets.