# Enhancing Knowledge Graph Question Answering through Classified Path Retrieval and Deductive Reasoning

1st Hao Wu
*Shanghai University*,
Shanghai, China
planet@shu.edu.cn

2nd Xiangfeng Luo*
*Shanghai University*,
Shanghai, China
luoxf@shu.edu.cn

3rd Jianqi Gao
*Shanghai University*,
Shanghai, China
jianqi_gao@shu.edu.cn

4th Dian Huang
*Shanghai University*,
Shanghai, China
dhuang99101@shu.edu.cn

***Abstract*—Knowledge base question answering (KBQA) aims to retrieve precise answers from knowledge graphs using natural language queries. While large language models (LLMs) enhance KBQA with strong reasoning capacities, they still face challenges such as noisy retrieval paths and inherent difficulties in reasoning over graph structures. To address these issues, we propose a novel framework, Path Retrieval and Deductive Reasoning (PRDR), which reformulates path retrieval as a binary classification task to suppress irrelevant paths and facilitate accurate reasoning. Retrieved paths are converted into propositional forms, enabling LLMs to perform structured deductive inference. Our method achieves state-of-the-art results on the WebQuestionsSP and ComplexWebQuestions benchmarks, demonstrating robust performance in complex multi-hop question answering.**

***Index Terms*—Knowledge base question answering, Large language models, Path retrieval, Deductive reasoning**

## I. Introduction

Knowledge base question answering (KBQA) aims to answer natural language questions by leveraging structured knowledge from large-scale knowledge graphs (KGs) [1]. As information grows exponentially, KBQA has become increasingly critical across various domains due to its ability to provide precise and efficient information access [2]. By bridging natural language queries with structured knowledge representations, KBQA systems allow users to interact with complex knowledge through intuitive interfaces.

The emergence of large language models (LLMs) has significantly advanced KBQA. These models exhibit strong capabilities in natural language understanding and generation, enabling enhanced performance on complex QA tasks [3]. Traditional KBQA methods often rely on semantic parsing and information retrieval, which struggle with language variation and multi-step reasoning. In contrast, LLM-based approaches demonstrate better semantic understanding and response generation capabilities, leading to notable performance gains [4].

Nevertheless, integrating LLMs into KBQA remains challenging due to two persistent issues. First, retrieving relevant paths from KGs often introduces substantial noise [5]. Large-scale KGs contain numerous interconnected entities, resulting in multiple candidate paths, many of which are irrelevant or misleading. This noise not only increases computational cost but also misleads the reasoning process, degrading answer accuracy. Second, LLMs inherently lack effective reasoning capabilities over graph-structured data. Although proficient in sequential text processing, they struggle to capture complex relational patterns and multi-hop dependencies within KGs [6], [7], limiting their applicability in structured reasoning tasks.

To overcome these challenges, we propose a novel approach that reinvents path retrieval and reasoning for KBQA. Our method introduces three key innovations:

- We reformulate path retrieval as a binary classification task, associating natural language questions with relevant KG relations to reduce noise and improve retrieval precision.
- We introduce a structured deductive reasoning framework that represents retrieved paths as propositional forms (premises and conclusions), fully leveraging the innate logical abilities of LLMs.
- We design a joint learning strategy that integrates classification-based retrieval with deductive inference, enabling more accurate and interpretable answer derivation.

Extensive experiments demonstrate that our approach achieves state-of-the-art performance on widely-used benchmarks including WebQuestionsSP (WebQSP) [8] and ComplexWebQuestions (CWQ) [9], validating its effectiveness for both simple and complex questions.

## II. Related Work

### A. KBQA

Knowledge base question answering (KBQA) leverages structured knowledge bases to provide more accurate answers than traditional QA systems. Existing approaches are broadly categorized into two paradigms: information retrieval-based (IR-based) and semantic parsing-based (SP-based) [6]. IR-based methods retrieve relevant entities via semantic matching and path ranking, yet exhibit limitations in handling multi-hop reasoning. SP-based methods convert natural language questions into logical forms by leveraging pre-trained language models to predict operators and compose subqueries [10]. Recent progress in LLMs has significantly advanced KBQA, enabling few-shot [11] and zero-shot [12] learning. Emerging

* Xiangfeng Luo is the corresponding author.

research further investigates the use of LLMs' intrinsic reasoning abilities without external knowledge base interactions [13], extending the scope of conventional KBQA.

### B. Information Retrieval and Knowledge Reasoning

Information retrieval-based KBQA (IR-KBQA) answers natural language queries through structured retrieval and reasoning over knowledge bases. The framework operates through two core stages: information retrieval and knowledge reasoning. The retrieval phase identifies semantically aligned triples or relational paths from the knowledge graph, serving as a critical filter for noise reduction and quality assurance [14]. The reasoning phase interprets these paths to derive answers. While traditional graph-based methods (e.g., GNNs) approximate multi-hop reasoning, they often struggle with complex logical structures and noise contamination. Recent LLM-enhanced approaches (e.g., ToG, RoG) integrate large language models with KBs for iterative path exploration, significantly improving interpretability and multi-hop reasoning capability [15]. However, noisy retrieval paths remain a fundamental challenge, as irrelevant paths significantly compromise reasoning accuracy [16]. Emerging hybrid methods that combine neural retrieval with symbolic verification show promise in addressing this limitation [2].

### C. Text Classification in KBQA

Text classification plays a crucial role in KBQA [17], [18], enabling accurate query understanding and effective identification of relevant entities and relations in knowledge graphs. This approach classifies natural language queries into predefined categories [19], which are subsequently mapped to corresponding knowledge graph elements [20], thereby facilitating targeted information retrieval. Recent advances leverage transformer-based architectures like BERT [21], which have demonstrated strong performance across NLP tasks. When fine-tuned on KBQA-specific datasets, these models significantly improve the precision of path retrieval and reasoning processes [22], [23]. By enhancing the relevance of retrieved paths, text classification methods directly address key challenges including noisy retrieval paths and limited reasoning capabilities in current systems.

## III. Methodology

### A. Definition of the KBQA task

A knowledge graph (KG) is a structured representation of knowledge, typically represented as a set of triples:

$$G = \{(s, r, o) \mid s, o \in E, r \in R\},$$

where $\langle s, r, o\rangle$ denotes a relation $r$ between a subject $s$ and an object $o$, with $E$ representing the set of entities and $R$ is the set of relations in the graph. In the context of KBQA, the task can be formally defined as follows:

Let $Q = \{q_1, q_2, \ldots, q_n\}$ be the set of questions, $\mathcal{P} = \{\rho_1, \rho_2, \ldots, \rho_m\}$ be the set of relations connecting two entities or an entity with a literal. $A$ is the set of all possible answers related to question q. Given a knowledge base $\mathcal{K}$ and a natural language question $q \in Q$, the goal of KBQA is to return the correct answer $a \in A$, which corresponds to the query $q$.

### B. Path Retrieval

*a) Relation Path Representation:* In KBQA systems, the representation of relation paths is critical for effective reasoning. A path in the knowledge graph is defined as $\mathcal{P} = (\rho_1, \rho_2, \ldots, \rho_m)$, where each $\rho_i \in \mathcal{P}$ represents a relation, and $\mathcal{P}$ denotes the set of all possible relations in the graph.

We evaluate the relevance of a relation path $p$ to a user query $q$, we use a structured template. For each query-path pair, we create a unified textual representation that concatenates the natural language query $q$ with the sequential relation path $\rho_1, \ldots, \rho_m$. This unified format jointly captures the user's information need and the structural semantics of the candidate path, providing a solid foundation for subsequent relevance assessment.

*b) Sample Construction for Classifier Training:* The binary classifier requires high-quality supervised data to distinguish relevant from irrelevant relation paths. We construct the training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ as follows:

**Positive samples** ($y = 1$): For each training question $q$ with its ground-truth answer entity $a$, we extract all relation paths from the linked question entities to $a$ within the knowledge graph. Each resulting query-path pair forms a positive instance.

**Negative samples** ($y = 0$): To build a robust classifier, negative instances are generated through random sampling from the set of KG relations that are *not* associated with the correct answer for the given query. Specifically, we sample relations that appear in other, semantically distinct questions from the training corpus to ensure diversity and avoid trivial negatives. The ratio of positive to negative samples is maintained at 1:3, a configuration empirically validated to balance the classifier's precision and recall while mitigating potential bias toward the majority class.

Formally, for a question $q$ with gold answer $a$, let $\mathcal{R}_a$ be the set of relations on any valid reasoning path from the question entities to $a$. The positive set is $\mathcal{D}^+ = \{(q, r) \mid r \in \mathcal{R}_a\}$. The negative set is $\mathcal{D}^- = \{(q, r') \mid r' \in \mathcal{R} \setminus \mathcal{R}_a\}$, where $\mathcal{R}$ is the full relation vocabulary and $r'$ is sampled uniformly. The final labeled dataset is $\mathcal{D} = \mathcal{D}^+ \cup \mathcal{D}^-$.

*c) Model Training:* The unified representation of each query-path pair $\mathbf{x}_i$ serves as input to a pre-trained language model, specifically $\text{BERT}_{\text{Large}}$ [21]. The model is fine-tuned for the binary classification task defined by the dataset $\mathcal{D}$ constructed above.

The pre-trained language model $\text{PLM}(\cdot; \theta)$, parameterized by $\theta$, generates contextualized token representations for the input sequence $\mathbf{x}_i$: $\mathbf{h}(\mathbf{x}_i; \theta) = [\mathbf{h}_{\text{CLS}}, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_m]$. The $\mathbf{h}_{\text{CLS}}$ token embodies the aggregated semantic representation of the entire input. For classification, this representation is passed through a softmax layer:

$$f(\mathbf{h}_{\text{CLS}} \mid \mathbf{W}_s) = \sigma(\mathbf{W}_s \cdot \mathbf{h}_{\text{CLS}}), \tag{1}$$

where $\mathbf{W}_s$ is a trainable weight matrix and $\sigma$ is the sigmoid function. The model is optimized by minimizing the binary cross-entropy loss:

$$\mathcal{L}_{\text{class}}(\theta \cup \mathbf{W}_s; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \ell_{\text{ce}} \left( f(\mathbf{h}(\mathbf{x}_i; \theta) \mid \mathbf{W}_s), y_i \right), \tag{2}$$

where $\ell_{\text{ce}}(\cdot, \cdot)$ denotes the cross-entropy loss. Minimizing this objective improves the classifier's accuracy in identifying query-relevant relation paths, which is essential for filtering noisy candidates in the retrieval stage.

### C. Deductive Reasoning

Deductive reasoning enables the derivation of logically certain conclusions from verified premises. While large language models (LLMs) possess strong capabilities in textual inference, their ability to perform structured reasoning over graph-based knowledge remains limited. To bridge this gap, we propose a deductive framework that converts retrieved relation paths into propositional forms, thereby leveraging LLMs' inherent strength in factual and logical deduction within a structured context.

*a) Propositional Transformation of Paths:* Each retrieved relation path is transformed into a chain of natural-language propositions. Given a path $\mathcal{P} = (e_0 \xrightarrow{\rho_1} e_1 \xrightarrow{\rho_2} \dots \xrightarrow{\rho_m} e_m)$, it is converted into a sequence of declarative statements of the form "$e_{i-1}$ [relation $\rho_i$] $e_i$," where [relation $\rho_i$] is verbalized into a natural phrase (e.g., `studied_by` becomes "is studied by"). For multi-hop paths ($m > 1$), these statements are chained as connected premises, preserving the logical flow from the subject entity to the answer candidate. All candidate paths that pass the classifier filter (Section III-B) are transformed simultaneously. The resulting set of propositional chains is presented to the LLM within a structured prompt, instructing it to evaluate logical consistency and select the chain that best supports answering the original question.

*b) Handling of Multiple and Conflicting Paths:* The binary classifier serves as a high-precision filter, eliminating the majority of spurious or irrelevant paths. The remaining candidate paths—typically a small, high-confidence set—are all propositionalized and submitted to the LLM. The prompt explicitly asks the model to reason over all provided chains, compare their coherence with the question, and identify the most logically supported answer. This approach allows the LLM to resolve mild conflicts or ambiguities by leveraging its broader contextual knowledge, while the classifier ensures that blatantly incorrect paths are removed beforehand.

*c) Integrated Workflow:* As formalized in Algorithm 1, our method operates through three integrated phases: (1) entity linking identifies relevant entities in $q$; (2) the binary path classifier filters retrieved paths—a path $p$ is retained if its predicted relevance score $\mathbf{s}_1$ exceeds threshold $\tau_1$, or if its predicted irrelevance score $\mathbf{s}_2$ is below threshold $\tau_2$ (indicating low confidence in irrelevance); (3) deductive reasoning transforms the retained paths into propositional chains, which the LLM evaluates to identify the logically valid answer. This structured approach enhances reasoning reliability by combining precise, classification-guided retrieval with formalized logical evaluation.

**Algorithm 1** Path Retrieval and Deductive Reasoning (PRDR)

**Require:** Knowledge graph $\mathcal{G} = (E, R, T)$; Question $q$; Thresholds $\tau_1$, $\tau_2$
**Ensure:** Answer entity $a$

$\mathcal{E} \leftarrow \text{EntityLinking}(q, \mathcal{G})$
$\mathcal{P} \leftarrow \text{RetrievePaths}(\mathcal{E}, \mathcal{G}, 3)$
$\mathcal{P}^* \leftarrow \emptyset$
**for** each path $p \in \mathcal{P}$ **do**
  $\mathbf{x} \leftarrow \text{Encode}(q, p)$
  $\mathbf{h} \leftarrow \text{BERT}(\mathbf{x})$
  $\mathbf{s} \leftarrow \text{Softmax}(\mathbf{W}\mathbf{h}_{\text{CLS}})$
  **if** $\mathbf{s}_1 > \tau_1$ **or** $(\mathbf{s}_2 < \tau_2)$ **then**
    $\mathcal{P}^* \leftarrow \mathcal{P}^* \cup \{p\}$
  **end if**
**end for**
$\mathcal{S} \leftarrow \text{ConvertToPropositions}(\mathcal{P}^*)$
$i^* \leftarrow \text{LLM}(\text{BuildPrompt}(q, \mathcal{S}))$
$a \leftarrow \text{ExtractAnswer}(\mathcal{P}^*[i^*])$
**return** $a$

## IV. Experiments

In this section, we provide a detailed description of the dataset and experimental setup. We present the results of our main experiments, compare them with the baseline model, and discuss specific experimental cases to demonstrate the validity of our approach.

### A. Experiment Settings

*a) Datasets and Baselines:* We evaluate our method using two widely-recognized benchmark datasets for KBQA: WebQSP [8] and CWQ [9], both designed for reasoning over the Freebase knowledge base [24]. For consistency, we adopt the same training and testing splits as those used in previous research [25].

We compare PRDR with several KBQA baselines, which are categorized into three paradigms: Non-LLM methods, prompting-based LLM methods, and methods combining LLMs with Knowledge Graphs (KGs). In the inference path retrieval process, we use $\text{BERT}_{\text{Large}}$ [21] as a text classification model to identify and retain relevant and correct inference paths while filtering out relationships that are not relevant to each specific problem, and then we use LLM GPT-3.5-turbo for path inference.

*b) Training Details:* We use F1 score and Hits@1 as evaluation metrics, following standard KBQA practice. The F1 score measures the overall coverage of correct answers, while Hits@1 assesses the accuracy of the top-ranked answer. Additionally, for evaluating retrieval performance, we compute top-k accuracy to assess whether the correct relation paths appear within the top-k retrieved answers. For the classification model, we utilized the $\text{BERT}_{\text{Large}}$ model, trained for 3 epochs with a batch size of 32 and a learning rate of $1 \times 10^{-5}$. The maximum sequence length was set to 128. All experiments were conducted on a single NVIDIA A6000 GPU.

TABLE I
PERFORMANCE COMPARISON ON WEBQSP AND CWQ DATASETS

| Type | Methods | WebQSP | | CWQ | |
|---|---|---|---|---|---|
| | | Hits@1 | F1 | Hits@1 | F1 |
| Non-LLMs | KV-Mem [26] | 46.7 | 34.5 | 21.1 | 15.7 |
| | GraftNet [27] | 66.4 | 60.4 | 36.8 | 32.7 |
| | NSM [28] | 74.3 | 67.4 | 48.8 | 44.0 |
| | SR+NSM [29] | 68.9 | 64.1 | 50.2 | 47.1 |
| | SR+NSM+E2E [29] | 69.5 | 64.1 | 49.3 | 46.3 |
| | QGG [28] | 73.0 | 73.8 | 36.9 | 37.4 |
| LLMs Only | Alpaca-7B [30] | 51.8 | - | 27.4 | - |
| | LLaMA2-Chat-7B [31] | 64.4 | - | 34.6 | - |
| | Zero-shot (gpt-3.5) | 54.4 | 52.3 | 34.9 | 28.3 |
| | Few-shot (gpt-3.5) | 56.3 | 53.1 | 38.5 | 33.9 |
| | CoT (gpt-3.5) | 57.4 | 54.7 | 43.2 | 35.9 |
| LLMs + KGs | KD-CoT [32] | 68.6 | 52.5 | 55.7 | - |
| | ToG [33] | 75.1 | 72.3 | 57.6 | 57.0 |
| | UniKGQA [34] | 77.2 | 72.2 | 51.2 | 49.1 |
| | RoG [3] | 83.2 | 69.8 | 61.4 | 56.2 |
| | **PRDR** (gpt-3.5) | **83.6** | **77.0** | **66.4** | **62.8** |

### B. Main Results

Table I presents the performance comparison of PRDR against existing baselines on WebQSP and CWQ. With GPT-3.5 as the reasoning engine, PRDR consistently outperforms all competing methods across both evaluation metrics. Specifically, our approach achieves 83.6% Hits@1 and 77.0% F1 score on WebQSP, surpassing the previous state-of-the-art method RoG [3] by 0.4% and 7.2%, respectively. On the more challenging CWQ dataset, which contains complex multi-hop questions, PRDR obtains 66.4% Hits@1 and 62.8% F1, exceeding RoG by 5.0% and 6.6%.

These gains are particularly noteworthy when compared to other LLM+KG approaches such as ToG [33] and UniKGQA [34], where PRDR achieves absolute F1 improvements of 6.4% and 4.8% on WebQSP, and 15.2% and 13.7% on CWQ, respectively. The larger margin on CWQ highlights the effectiveness of our path classification filter in eliminating spurious reasoning chains, which are more prevalent in complex queries, and the propositional deductive reasoning stage in faithfully deriving correct answers.

### C. Comparison of Retrieval Models

We evaluate three retrieval models—BM25, Sentence-BERT, and $\text{BERT}_{\text{Large}}$—on the WebQSP and CWQ benchmarks. As shown in Fig. 1, $\text{BERT}_{\text{Large}}$ achieves the highest retrieval accuracy across all settings, demonstrating superior precision in identifying exact path matches. Sentence-BERT exhibits competitive performance but lags behind $\text{BERT}_{\text{Large}}$, particularly on the more complex CWQ dataset. BM25 shows adequate results on WebQSP but performs poorly on CWQ.

The performance gap is quantified using the Top-K accuracy metric:

$$\text{Acc@}K = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}\left(\text{Correct}(q, K)\right) \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $\text{Correct}(q, K)$ evaluates to 1 if the gold relation path for question $q$ appears within the top-$K$ retrieved candidates, and 0 otherwise.

$\text{BERT}_{\text{Large}}$'s lower partial retrieval accuracy further confirms its precise matching capability, whereas the higher partial accuracy of Sentence-BERT and BM25 suggests weaker discriminative power. These results establish $\text{BERT}_{\text{Large}}$ as the most robust retrieval model for KBQA tasks.

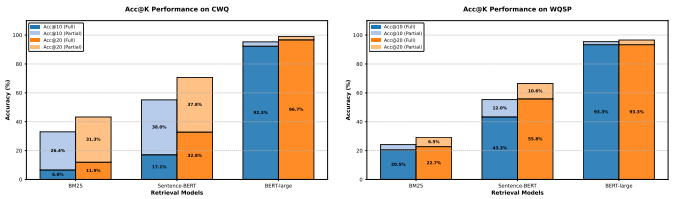



Fig. 1. Comparative evaluation of retrieval models on (a) CWQ and (b) WebQSP datasets. Error bars represent standard deviation over 5 runs.

### D. Threshold Selection and Sensitivity Analysis

The thresholds $\tau_1$ (relevance confidence) and $\tau_2$ (irrelevance confidence) in Algorithm 1 are critical for balancing precision and recall during the path-filtering stage. To determine their optimal values, we conduct a comprehensive grid search over $\tau_1 \in [0.5, 0.9]$ and $\tau_2 \in [0.5, 0.9]$ with a step size of 0.1, evaluating the F1 score on the validation splits of both WebQSP and CWQ benchmarks using GPT-3.5 as the reasoning engine.

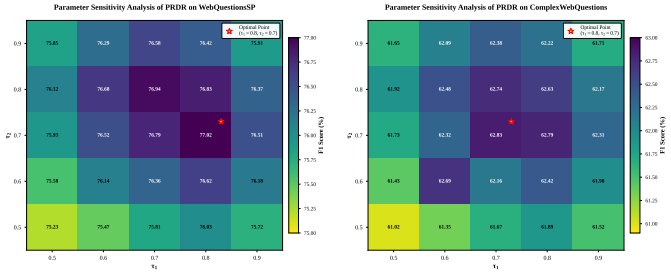
As shown in Fig. 2, our sensitivity analysis reveals consistent performance patterns across both datasets. For WebQSP (Fig. 2a), the highest F1 score of 77.02% is achieved at $\tau_1 = 0.8$ and $\tau_2 = 0.7$. Similarly, for CWQ (Fig. 2b), the optimal configuration occurs at the same parameter values, yielding an F1 score of 62.79%. The performance landscapes demonstrate that PRDR maintains stable performance across a wide range of threshold combinations, with F1 score less than 1.5% across all tested parameter pairs. This stability indicates that the classifier's confidence scores are well-calibrated and the filtering mechanism exhibits robustness to minor threshold perturbations.

### E. Case Study

The case studies in Table II illustrate that the proposed PRDR method enhances LLMs by integrating path retrieval and deductive reasoning to give correct answers. Although path retrieval alone can identify complete paths through binary classification, it may produce incorrect results due to the lack of logical reasoning. By incorporating deductive reasoning, the proposed method ensures accuracy of reasoning through a structured thought process, effectively filters irrelevant paths, and returns correct results. Unlike traditional approaches relying on sequential reasoning such as CoT, PRDR performs path retrieval by categorizing tasks, selecting the most relevant reasoning paths and refining them with deductive logic. This approach greatly enhances the ability of LLMs to handle KBQA tasks. The experimental results demonstrate that when querying "*What is the name of the sports team that plays in Old Trafford, Greater Manchester?*", the CoT approach yielded an incorrect reasoning path and consequently produced an erroneous answer due to its inability to detect and rectify

TABLE II
CASE STUDY OF PRDR

| | |
|---|---|
| **Question** | *What is the name of the sports team that plays in Old Trafford, Greater Manchester?* |
| **Gold Answer** | *Newton Heath L&YR F.C.* |
| **CoT Prompt** | You're an expert in Knowledge Graph Question Answering. Now I have a question: (What is the name of the sports team that plays in Old Trafford, Greater Manchester?), and a series of potential reasoning path that lead to its answer:<br>0: The reasoning path is: Old Trafford, Greater Manchester → sports.sports_team.location → Manchester United F.C. → sports.sports_team.arena_stadium → Old Trafford, its conclusion is: The name of the sports team that plays in Old Trafford, Greater Manchester is Manchester United F.C.<br>1: The reasoning path is: Old Trafford, Greater Manchester → sports.sports_team.location → Manchester United F.C. → sports.sports_team.previously_known_as → Newton Heath L&YR F.C., its conclusion is: The name of the sports team that plays in Old Trafford, Greater Manchester is Newton Heath L&YR F.C.<br>2: The reasoning path is: Old Trafford, Greater Manchester → location.location.contains → Greater Manchester → travel.travel_destination.tourist_attractions → Great Manchester Run, Museum of Science and Industry, Pennington Flash Country Park, its conclusion is: The sports team that plays in Old Trafford, Greater Manchester is not directly identified by the tourist attractions in Greater Manchester.<br>You are asked to output the correct reasoning path after thinking step-by-step and verifying the final answer for the given question. |
| **Output:** | The correct reasoning path is Path 0<br>The correct answer is: Manchester United F.C. |
| **PRDR (ours)** | You're an expert in Knowledge Graph Question Answering. Now I have a question: (What is the name of the sports team that plays in Old Trafford, Greater Manchester?), and a series of potential reasoning path that lead to its answer:<br>0: The reasoning path is: Old Trafford, Greater Manchester → sports.sports_team.location → Manchester United F.C. → sports.sports_team.arena_stadium → Old Trafford, its conclusion is: The name of the sports team that plays in Old Trafford, Greater Manchester is Manchester United F.C.<br>1: The reasoning path is: Old Trafford, Greater Manchester → sports.sports_team.location → Manchester United F.C. → sports.sports_team.previously_known_as → Newton Heath L&YR F.C., its conclusion is: The name of the sports team that plays in Old Trafford, Greater Manchester is Newton Heath L&YR F.C.<br>2: The reasoning path is: Old Trafford, Greater Manchester → location.location.contains → Greater Manchester → travel.travel_destination.tourist_attractions → Great Manchester Run, Museum of Science and Industry, Pennington Flash Country Park, its conclusion is: The sports team that plays in Old Trafford, Greater Manchester is not directly identified by the tourist attractions in Greater Manchester.<br>You are tasked to verify the correct reasoning path for the given question by deductiving reasoning. If a correct path exists, return the index of the most likely correct path. Otherwise, return 'None'. |
| **Output (ours):** | The most likely correct path is **Path 1**<br>Final Answer: **Newton Heath L&YR F.C.** |




(a) WebQSP dataset (b) CWQ dataset

Fig. 2. Parameter sensitivity analysis of PRDR thresholds on WebQSP and CWQ validation sets. The heatmaps visualize F1 scores across different combinations of $\tau_1$ and $\tau_2$. The gold star marks the optimal configuration. Darker colors indicate higher F1 scores.

reasoning errors. In contrast, our PRDR method successfully identified and validated the correct reasoning path (**Path 1**), ultimately confirming the ground truth entity (**Newton Heath L&YR F.C.**) as the accurate result.

### *F. Ablation Study*

TAblation experiments with GPT-3.5 evaluate the contribution of each core module. As shown in Table III, removing either the path retrieval or deductive reasoning component results in significant performance degradation on both WebQSP and CWQ benchmarks. Disabling path retrieval leads to an F1 drop of 3.52 and 2.84 percentage points on WebQSP and CWQ, respectively, underscoring its role as the primary filter for noisy reasoning paths. Removing deductive reasoning causes reductions of 2.76 and 2.19 percentage points, confirming that structured logical inference adds substantial value beyond simple path extraction.

The full PRDR framework achieves the highest scores, demonstrating that the synergistic integration of classification-based retrieval and propositional reasoning is essential for optimal KBQA performance. The consistent trend across datasets validates the robustness of this design.

TABLE III
ABLATION STUDY OF PRDR COMPONENTS ON WEBQSP AND CWQ DATASETS

| Variant | WebQSP (F1) | CWQ (F1) |
|---|---|---|
| Full PRDR | **77.01** | **62.82** |
| w/o Path Retrieval | 73.49 (-3.52) | 59.98 (-2.84) |
| w/o Deductive Reasoning | 74.25 (-2.76) | 60.63 (-2.19) |

## V. Conclusion

We proposed PRDR, a novel method that enhances LLM reasoning on knowledge graphs for KBQA. Specifically, our approach transforms path retrieval into a classification task, enabling the retrieval of reasoning paths relevant to the question while effectively filtering out a large number of irrelevant paths. During the reasoning phase, PRDR enables LLMs to perform faithful reasoning on the retrieved paths in a deductive manner, thereby obtaining accurate answers. Experimental results demonstrate that PRDR significantly outperforms existing baseline methods on the WebQSP and CWQ benchmarks, achieving state-of-the-art performance.

## References


[1] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[2] Yuan Sui, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi, "Fidelis: Faithful reasoning in large language model for knowledge graph question answering," *arXiv preprint arXiv:2405.13873*, 2024.

[3] Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan, "Reasoning on graphs: Faithful and interpretable large language model reasoning," *arXiv preprint arXiv:2310.01061*, 2023.

[4] Haoran Luo, Zichen Tang, Shiyao Peng, Yikai Guo, Wentai Zhang, Chenghao Ma, Guanting Dong, Meina Song, Wei Lin, et al., "Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models," *arXiv preprint arXiv:2310.08975*, 2023.

[5] Hang Zhang, Yeyun Gong, Xingwei He, Dayiheng Liu, Daya Guo, Jiancheng Lv, and Jian Guo, "Noisy pair corrector for dense retrieval," *arXiv preprint arXiv:2311.03798*, 2023.

[6] Feng Gao, Yan Yang, Peng Gao, Ming Gu, Shangqing Zhao, Yuefeng Chen, Hao Yuan, Man Lan, Aimin Zhou, and Liang He, "Self-supervised bgp-graph reasoning enhanced complex kbqa via sparql generation," *Information Processing & Management*, vol. 61, no. 5, pp. 103802, 2024.

[7] Costas Mavromatis and George Karypis, "Gnn-rag: Graph neural retrieval for efficient large language model reasoning on knowledge graphs," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 16682–16699.

[8] Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina 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)*, 2016, pp. 201–206.

[9] Alon Talmor and Jonathan Berant, "The web as a knowledge-base for answering complex questions," *arXiv preprint arXiv:1803.06643*, 2018.

[10] Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Manzil Zaheer, Hannaneh Hajishirzi, Robin Jia, and Andrew McCallum, "Knowledge base question answering by case-based reasoning over subgraphs," in *International conference on machine learning*. PMLR, 2022, pp. 4777–4793.

[11] Mayur Patidar, Riya Sawhney, Avinash Singh, Biswajit Chatterjee, Indrajit Bhattacharya, et al., "Few-shot transfer learning for knowledge base question answering: Fusing supervised models with in-context learning," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 9147–9165.

[12] Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang, "Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 18608–18616.

[13] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi, "Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family," in *International Semantic Web Conference*. Springer, 2023, pp. 348–367.

[14] Muhammad Ahmad, "Toward a unified framework for information retrieval in large language model applications: Balancing textual and graph-based knowledge sources," 2025.

[15] Zeping Yu, Yonatan Belinkov, and Sophia Ananiadou, "Back attention: Understanding and enhancing multi-hop reasoning in large language models," *arXiv preprint arXiv:2502.10835*, 2025.

[16] Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, and Xiao Huang, "A survey of graph retrieval-augmented generation for customized large language models," *arXiv preprint arXiv:2501.13958*, 2025.

[17] Zhipeng Liu, Jing He, Tao Gong, Heng Weng, Fu Lee Wang, Hai Liu, and Tianyong Hao, "Improving topic tracing with a textual reader for conversational knowledge based question answering," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[18] Huiying Li, Yuxi Feng, and Liheng Liu, "Fusing essential text for question answering over incomplete knowledge base," *Electronics*, vol. 14, no. 1, pp. 161, 2025.

[19] Yaohui Wang, Yuxian Zhao, Qiang Hao, Jian Zhang, Xiaohu Sun, and Xueqiang Lyu, "Question-and-answer intention classification in the coal mine production field based on prompt learning," in *Proceedings of the 5th International Conference on Computer Information and Big Data Applications*, 2024, pp. 992–998.

[20] Kamal Taha, Paul D Yoo, Chan Yeun, and Aya Taha, "Text classification: A review, empirical, and experimental evaluation," *arXiv preprint arXiv:2401.12982*, 2024.

[21] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*. Minneapolis, Minnesota, 2019, vol. 1, p. 2.

[22] Sumit Sharma, Sarika Jain, Mukesh Kumar Tiwari, and Sohan lal, "Classifying the state of knowledge-based question answering: patterns, progress, and prospects," *International Journal of Computers and Applications*, pp. 1–13, 2024.

[23] Yuan Sui, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi, "Fidelis: Faithful reasoning in large language models for knowledge graph question answering," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 8315–8330.

[24] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.

[25] Donghan Yu, Sheng Zhang, Patrick Ng, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Yiqun Hu, William Wang, Zhiguo Wang, and Bing Xiang, "Decaf: Joint decoding of answers and logical forms for question answering over knowledge bases," *arXiv preprint arXiv:2210.00063*, 2022.

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

[27] Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W Cohen, "Open domain question answering using early fusion of knowledge bases and text," *arXiv preprint arXiv:1809.00782*, 2018.

[28] Gaole He, Yunshi Lan, Jing Jiang, Wayne Xin Zhao, and Ji-Rong 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*, 2021, pp. 553–561.

[29] Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen, "Subgraph retrieval enhanced model for multi-hop knowledge base question answering," *arXiv preprint arXiv:2202.13296*, 2022.

[30] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto, "Stanford alpaca: An instruction-following llama model," 2023.

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[32] Jinheon Baek, Alham Fikri Aji, and Amir Saffari, "Knowledge-augmented language model prompting for zero-shot knowledge graph question answering," *arXiv preprint arXiv:2306.04136*, 2023.

[33] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo, "Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph," *arXiv preprint arXiv:2307.07697*, 2023.

[34] Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen, "Unikgqa: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph," *arXiv preprint arXiv:2212.00959*, 2022.