Diagnosing and Remedying Representation Deficiencies for Deterministic Reasoning in KGQA

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated increasingly strong reasoning capabilities, achieving remarkable progress in knowledge graph question answering (KGQA). However, a key challenge in such systems is non-deterministic reasoning, where the model indecisively activates multiple semantically related knowledge graph edges for a given query, frequently leading to incorrect answers. To address this issue, we propose **D**iagnosing and **R**emedying Representation Deficiencies for **D**eterministic **R**easoning in KGQA (DR^2). DR^2 identifies and localizes non-deterministic reasoning behaviors, uncovering the underlying semantic representation deficiencies in LLMs. Building on this diagnosis, we design abductive reasoning-based preference learning, which promotes fine-grained semantic discrimination and mitigates non-deterministic reasoning errors. Experimental results demonstrate that the proposed DR^2 significantly outperforms several strong baselines, achieving state-of-the-art performance on the widely used WebQSP and CWQ benchmarks.[1]

1 Introduction

Large language models have demonstrated increasingly strong reasoning capabilities, leading to remarkable improvements in Knowledge Graph Question Answering (KGQA) tasks. Current approaches primarily leverage LLMs' powerful capabilities through two paradigms: in-context learning augmented with retrieved evidence to directly generate answers or reasoning chains (Li et al., 2023; Nie et al., 2024), and interactive, step-by-step frameworks where the model decomposes questions or traverses the knowledge graph through tool calls (Gu et al., 2023; Huang et al., 2024; Sun et al., 2024). To further enhance precision, supervised fine-tuning on questions and synthesized reasoning paths is widely used to enable more accurate and direct generation of search queries (LUO et al., 2024; Luo et al., 2024; Bu et al., 2025; Xu et al., 2025b). These methods advance KGQA by more effectively combining the knowledge of LLMs with the knowledge graphs, enhancing the reliability of reasoning process.

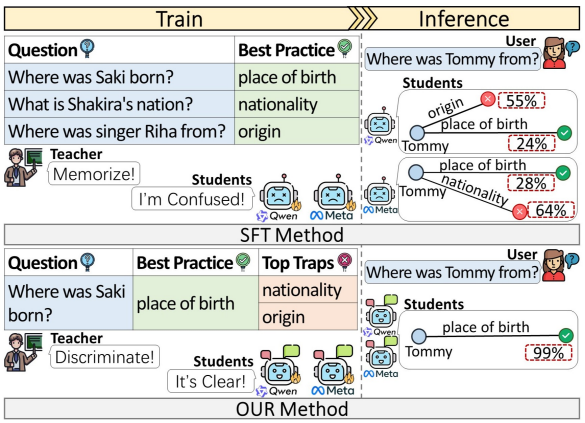

Figure 1: Comparing SFT's memorization-induced indecisiveness with our method's deterministic discrimination on semantically related edges.

However, a key challenge in such systems is non-deterministic reasoning, where the model indecisively activates multiple semantically related knowledge graph edges for a given query, frequently leading to incorrect answers. Existing Supervised Fine-Tuning (SFT) approaches in KGQA fundamentally fall short of addressing this semantic ambiguity. As Figure 1 illustrates, the indecisiveness of SOTA models on semantically similar knowledge graph edges—such as confusing "place of birth" with "nationality"—accounts for up to 53% of overall errors (see Figure 7). This critical failure stems from a key limitation of SFT: while it optimizes for matching the correct tool call during training, it does not equip the model with the ability to internally discriminate between closely related relations. We empirically

[1] Our codes and data will be released upon acceptance.

verify this in a controlled setting with the easily confusable tool calls "language spoken" and "official language". Although SFT method achieves high accuracy on the training set, the performance on unseen test data drops to 75%, as shown in Figure 8. This 25% accuracy gap reveals the model's reliance on memorization rather than true semantic discrimination, and its persistent vulnerability to non-deterministic reasoning when encountering unfamiliar examples.

To mitigate this issue, we propose **D**iagnosing and **R**emedying Representation Deficiencies for **D**eterministic **R**easoning in KGQA (**DR^2**), which enhances the model's ability to discriminate between semantically related knowledge graph edges. DR^2 identifies and localizes non-deterministic reasoning behaviors, uncovering the underlying semantic representation deficiencies in LLMs. Building on this diagnosis, we designs abductive reasoning-based preference learning, which promotes fine-grained semantic discrimination and mitigates non-deterministic reasoning errors. DR^2 enhances the model's capability to distinguish fine-grained semantic differences, our main contributions are:

- We propose a novel method DR^2 to detect non-deterministic reasoning and locate semantic representation deficiencies in LLMs.
- We propose using abductive reasoning method to acquire fine-grained semantic preference data to mitigate the deficiencies.
- The proposed **DR^2** method achieved state-of-the-art performance on two widely used benchmarks WebQSP and CWQ.

2 Related Works

Preferences Alignment. Recent advances in LLM alignment that leverage human preferences offer a promising direction for enhancing the discriminative capability of KBQA systems. The established method Reinforcement Learning from Human Feedback (RLHF) trains a reward model to optimize the policy model (Ouyang et al., 2022; Bai et al., 2022). Direct Preference Optimization (DPO) eliminates the separate reward model by training directly on output preference pairs (Rafailov et al., 2023). Subsequent work simplified DPO by removing the reference model (Meng et al., 2024; Hong et al., 2024). Most recently, IOPO improves DPO by constructing bidirectional preference pairs, improving the model's fine-grained discriminative capability (Zhang et al., 2025).

Knowledge Graph Question Answering. Early research can be categorized into rule-based methods, that leverage the symbolic structure of Knowledge Bases(KB) for retrieval or parsing (Sun et al., 2018; Zhang et al., 2022; Ye et al., 2022) and embedding-based methods which utilize neural networks to learn latent representations of entities and relations for reasoning (Miller et al., 2016; Yasunaga et al., 2021; Jiang et al., 2022). With the advent of large language models (LLMs), KBQA methods have evolved significantly. Current approaches predominantly leverage LLMs' powerful reasoning capabilities, primarily through in-context learning (ICL) augmented with retrieved evidence to generate answers or reasoning chains directly (Li et al., 2023; Nie et al., 2024), as well as through step-by-step frameworks where the model interactively decomposes questions or traverses the knowledge graph (Gu et al., 2023; Huang et al., 2024; Sun et al., 2024). To further enhance performance and alignment with specific KB schemas, supervised fine-tuning (SFT) on datasets of question-logical form pairs has been effectively adopted to enable more accurate and direct generation of structured queries (Luo et al., 2024; LUO et al., 2024; Jiang et al., 2025; Xu et al., 2025b; Bu et al., 2025).

While fine-tuning enhances the model's ability to understand questions and produce formally correct queries, the fine-tuned models may still lack robust discriminative capabilities, often leading to non-deterministic reasoning when faced with semantically similar edges. To mitigate this issue, we propose the Diagnosing and Remedying Representation Deficiencies for Deterministic Reasoning in KGQA to explicitly enhance the model's ability to discriminate between edges with close semantic meanings, thereby improving the precision and reliability in KBQA tasks.

3 Preliminary

In this section, we introduce several basic concepts and definitions used in our work.

Knowledge Graph(KG). A Knowledge Graph is a set of triples, denoted as $KG = \{(e_0, r, e_1) | e_0, e_1 \in \mathcal{E}, r \in \mathcal{R}\}$, where $\mathcal{E}$ is the set of entities and $\mathcal{R}$ is the set of relations.

Reasoning Path. Given a query, a reasoning path

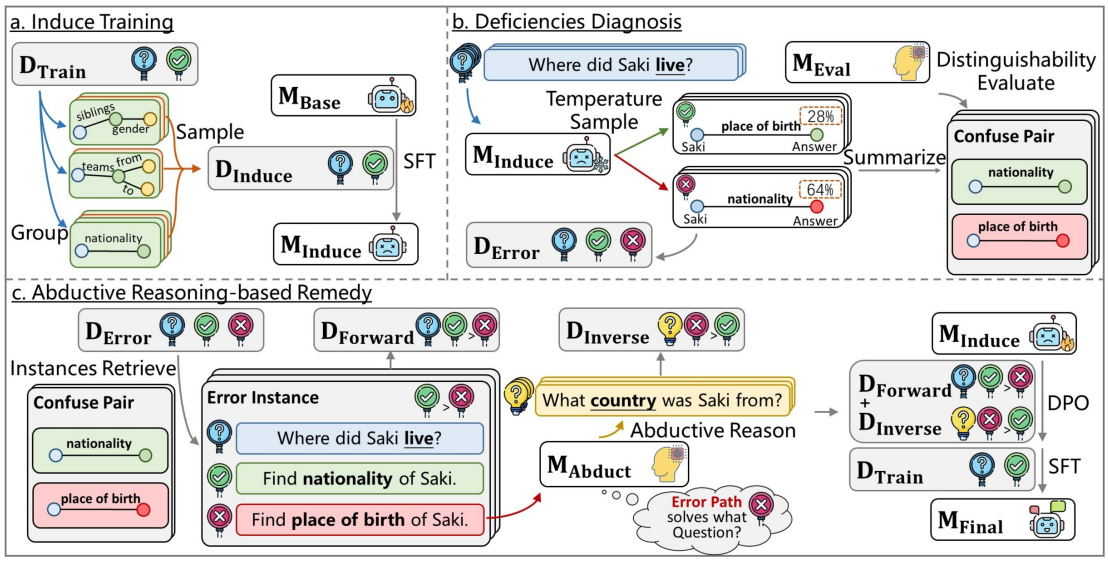

Figure 2: The overall framework of our DR^2. The diagnosing process trains an induced model with limited data (a) and diagnoses its deficiencies by analyzing the reasoning behaviors under temperature sampling (b). The remedying process leverages abductive reasoning to construct preference data for training the final model (c).

P is a sequence from the topic entity to the answer entity. It is formally represented as $P = e_{topic} \xrightarrow{r_1} e_1 \xrightarrow{r_2} ... \xrightarrow{r_n} e_{ans}$, where e_{topic} is the topic entity and e_{ans} is the answer entity. In the training set D_{Train}, the golden reasoning path P_{gold} is directly derived from its annotated SPARQL query.

Reasoning Structure. The reasoning structure is an abstraction of a reasoning path, obtained by removing all entity nodes and retaining only the sequence of relations. For a path $P = e_{topic} \xrightarrow{r_1} e_1 \xrightarrow{r_2} ... \xrightarrow{r_l} e_{ans}$, its reasoning structure is defined as $P^S = [r_1, r_2, ...r_l]$.

4 Approach

In this section, we provide a detailed description of our proposed DR^2 method. As illustrated in Figure 2, our approach to mitigating semantic representation deficiencies is decomposed into three tasks: Inducing Training, Deficiencies Diagnosis and Abductive Reasoning-based Remedy.

4.1 Inducing Training

Inducing Training aims to train an "induced model" to systematically expose the model's inherent semantic representation deficiencies. As shown in Figure 8, the model exhibits non-deterministic reasoning during early SFT, where the probabilities for the two similar paths rise synchronously. To induce this behavior, we design a training strategy based on reasoning structure classification.

Reasoning Structure Classification. For each data (Q, P_{Gold}) in the training set D_{Train}, we compute its reasoning structure P^S_{Gold}. We then partition D_{Train} into K mutually exclusive subsets $\{G_1, G_2, ...G_K\}$, where all data within the same subset share the same reasoning structure.

Non-determinism Inducing Training. To ensure the model is exposed to all reasoning structures, we randomly sample a small proportion of data from each subset G_i according to a pre-defined ratio to form the inducing training set D_{Induce}. We then fine-tune the base model M_{Base} using D_{Induce} to obtain the induced model M_{Induce}.

4.2 Deficiencies Diagnosis

After obtaining the induced model M_{Induce}, we are able to diagnose the model's semantic representation deficiencies by analyzing its non-deterministic reasoning. We first collect confusion errors from M_{Induce}, and summarize them into confusing structure pairs. A Semantic Distinguishability Evaluation is then proposed to pinpoint the pairs stemming from inherent semantic representation deficiencies, providing precise targets for remediation.

Confusion Error Collection. For each data (Q, P_{Gold}) in the training set D_{Train}, we feed the question Q into the induced model M_{Induce} and obtain m reasoning paths with a temperature sampling ($Sample_{Temp}$) process:

$$P_{Gen} = Sample_{Temp}(M_{Induce}, Q). \quad (1)$$

If the structure of the generated path differs from

that of the golden path, it forms a confusion error instance (Q, P_{Gold}, P_{Gen}). All such instances constitute the confusion error instance set D_{Error}.

To focus on systematic errors caused by semantic representation deficiencies, we summarize errors at the reasoning structure level. For each error instance (Q, P_{Gold}, P_{Gen}) in D_{Error}, we extract the reasoning structure of the golden path $S_1 = P^S_{Gold}$ and the generated path $S_2 = P^S_{Gen}$, then (S_1, S_2) forms an unordered confusion pair.

Semantic Distinguishability Evaluation. To pinpoint the confusion that arises from intrinsic deficiencies of the model, we evaluate the semantic distinguishability of each summarized confusion pair (S_1, S_2). We first construct two question groups G_1 and G_2 which contains all questions from D_{Train} whose golden reasoning structure is included in S_1 and S_2, respectively:

$$G_i = \{Q | (Q, P) \in D_{Train} \wedge P^S = S_i\}. \quad (2)$$

We randomly sample a question Q from $G_1 \cup G_2$, and then classify it into G_1 or G_2 using a general-purpose LLM M_{Eval}. The average classification accuracy (Acc) over multiple sampled questions yields the distinguishability score $DistSc$ using:

$$DistSc = Acc(M_{Eval}(Q) = Group(Q)). \quad (3)$$

A high $DistSc$ score indicates that the questions are semantically distinct, and the model's confusion stems from its representation deficiency. Pairs with scores above a predefined threshold form the target deficiencies set T_{Defect}, which is used for further remediation.

4.3 Abductive Reasoning-based Remedy

For each target deficiency in T_{Defect}, we extract the original error instances from D_{Error} as the forward preference data. We then leverage the abductive reasoning capability of a general-purpose LLM to construct inverse preference data. Finally, joint training on both forward and inverse preferences guides the model to deeply understand and discriminate between the confusion pairs, thereby remedying its representation deficiencies.

Error Instances Retrieval. For each target confusion pair in T_{Defect}, we filter all matching error instances (Q, P_{Gold}, P_{Gen}) from the error set D_{Error}. For each instance, a forward preference tuple $(Q, P_{Gold} \succ P_{Gen})$ is constructed, which is added to the forward preference set $D_{Forward}$.

Inverse Question Generation. For each matched error instance (Q, P_{Gold}, P_{Gen}), we employ a general-purpose LLM M_{Abduct} to perform abductive reasoning, inferring a new natural language question Q_{Inv} for which P_{Gen} becomes the uniquely correct reasoning path:

$$Q_{Inv} = M_{Abduct}(P_{Gen}). \quad (4)$$

This process yields a inverse preference tuple $(Q_{Inv}, P_{Gen} \succ P_{Gold})$, which is added to the inverse preference set $D_{Inverse}$

Preference Learning. The induced model M_{Induce} is optimized using Direct Preference Optimization on the union of the forward preference set $D_{Forward}$ and the inverse preference set $D_{Inverse}$. This directly targets the identified semantic representation deficiencies. The model is then Supervised Fine-Tuned on the original training set D_{Train} to produce the final model M_{Final}, which enhances its overall reasoning ability.

5 Experiment

In this section, we present a comprehensive experimental evaluation of the proposed DR^2 method. We first introduce the experimental setup, including the benchmarks, evaluation metrics, and baseline methods. Following this, we report the main results and a series of analytical experiments to examine the characteristics of DR^2 against baseline methods from multiple perspectives.

5.1 Experimental Settings

Benchmarks. To evaluate the performance of our proposed method on KGQA tasks, we employ two widely used benchmarks: WebQSP (Yih et al., 2016) and CWQ (Talmor and Berant, 2018).

Metrics. We adopt standard metrics for KGQA evaluation: Hits, F1 and Hits@1 following previous works. In our work, Hits measures if any predicted answer is correct, while Hits@1 specifically checks if the first predicted answer is correct. See detailed definitions in Appendix C.

Baselines. We compare DR^2 against previous state-of-the-art and representative KGQA methods, including strong baselines such as MemQ (Xu et al., 2025b) and KaeDe (Bu et al., 2025), as well as other notable methods for broader context.

Implementation Details. To ensure a fair comparison, we use Llama2-7B-chat (Touvron et al., 2023) as our base language model, consistent with the setup in recent methods like MemQ (Xu

Method	Society		Entertainment		Culture		Average	
	Hits@1	F1	Hits@1	F1	Hits@1	F1	Hits@1	F1
KaeDe	0.620	0.632	0.519	0.634	0.696	0.663	0.589	0.640
MemQ	0.741	0.773	0.561	0.648	0.696	0.790	0.648	0.718
DR^2	**0.759**	**0.808**	**0.603**	**0.677**	**0.732**	**0.798**	**0.681**	**0.745**

Table 1: Performance comparison of DR^2 and baselines on a confusing subset of WebQSP and CWQ.

et al., 2025b) and KaeDe (Bu et al., 2025). All our experiments are conducted on a machine with 4 NVIDIA RTX 4090 GPUs. During inference, we use a beam size of 4 for both datasets.

5.2 Main Result

Method	WebQSP		CWQ	
	Hits	F1	Hits	F1
KV-Mem (Miller et al., 2016)	0.467	0.345	0.184	0.157
GraftNet (Sun et al., 2018)	0.664	0.604	0.368	0.327
QGG (Lan and Jiang, 2020)	0.730	0.738	0.369	0.374
DECAF(Yu et al., 2022)	0.821	0.788	-	-
UniKGQA(Jiang et al., 2022)	0.751	0.702	0.507	0.480
ToG(Sun et al., 2024)	0.826	-	0.676	-
KG-Agent(Jiang et al., 2025)	0.833	0.810	0.722	0.692
FiDeLiS(Sui et al., 2025)	0.844	0.783	0.715	0.643
AMAR(Xu et al., 2025a)	0.843	0.812	0.831	0.785
FM-KBQA(Gao et al., 2025)	0.873	0.842	0.795	0.687
RoG(LUO et al., 2024)	0.857	0.708	0.626	0.562
ChatKBQA(Luo et al., 2024)	0.864	0.835	0.860	0.813
GGI-MAB(Tang et al., 2025)	0.866	0.756	0.794	0.720
EPERM(Long et al., 2025)	0.888	0.724	0.662	0.589
KaeDe(Bu et al., 2025)*	0.908	0.819	0.880	0.801
MemQ(Xu et al., 2025b)*	0.890	0.860	0.878	0.836
DR^2	**0.918**	**0.889**	**0.898**	**0.860**

Table 2: The results of our method DR^2 compared with previous approaches on WebQSP and CWQ. The asterisk * denotes the results we reproduced.

The main results of our proposed DR^2 on the WebQSP and CWQ datasets benchmarks are presented in Table 2 and Table 3. Our method achieves state-of-the-art performance across all key metrics on both benchmarks. This consistent and superior performance provides empirical evidence that our approach effectively mitigates the semantic representation deficiency in LLMs, confirming the effectiveness of diagnosing and remedying representation deficiencies in enhancing deterministic reasoning for KGQA. These results not only demonstrate the practical advantages of our method, but also highlight the fundamental importance of semantic representation robustness in complex reasoning tasks with LLMs.

Method	WebQSP	CWQ
RoG (LUO et al., 2024)	0.795	0.567
KaeDe (Bu et al., 2025)	0.830	0.798
MemQ (Xu et al., 2025b)	0.847	0.805
DR^2	**0.882**	**0.835**

Table 3: The results of our method DR^2 compared with previous approaches on WebQSP and CWQ using the strict Hits@1 metric (see Appendix C) for details.

5.3 Discrimination Capability Evaluation

To further evaluate the capability of DR^2 in discriminating between semantically related knowledge graph edges, we constructed a challenging subset derived from the WebQSP and CWQ test sets. This subset comprises 671 questions whose gold reasoning paths contain knowledge graph edges that are hard to distinguish. The questions are categorized into three domains (Society, Entertainment and Culture), based on their topic entities.

As shown in Table 1, DR^2 achieves consistent improvements over all baseline methods across all evaluation metrics on the challenging subset. These results confirm that our method effectively mitigates the semantic representation deficiency for semantically related edges.

5.4 Mitigation of Representation Deficiencies

To provide direct evidence of how DR^2 mitigates representation deficiencies, we compare its internal representations with those of MemQ on the identified defect set T_{Defect}. For each diagnosed defect pair (S_1, S_2), we input the corresponding questions and golden paths into the model to extract the hidden state of the last token, which forms representation groups G_1 and G_2. We then project the representations onto the first two principal components using Principal Component Analysis and visualize the distributions with Kernel Density Estimation. In addition, we calculate the Separability Score ($SepSc$) to quantify this difference, defined as the ratio of average inter-

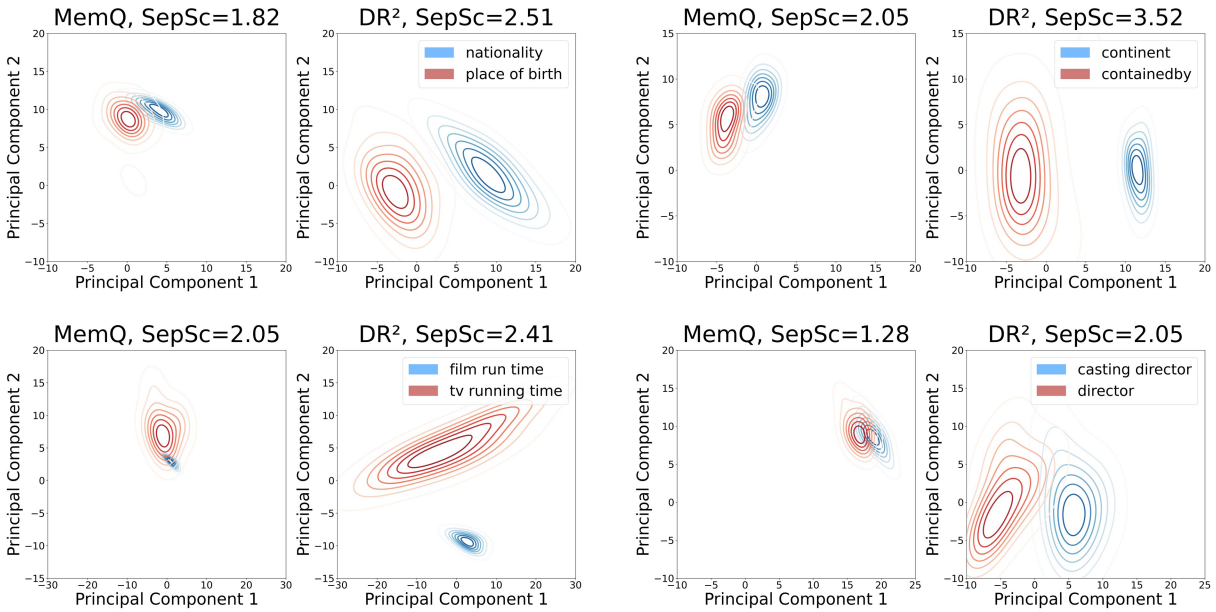

Figure 3: Comparing the separability of representations between DR^2 and MemQ.

group Euclidean distance (GD) to average intra-group Euclidean distance (D):

$$SepSc(G_1, G_2) = \frac{GD(G_1, G_2)}{\frac{1}{2} \times (D(G_1) + D(G_2))}. \quad (5)$$

As shown in Figure 3, compared to MemQ, DR^2 produces more separated representation clusters for the structures within each defect pair compared to MemQ, with a clearer boundary between them. Quantitatively, DR^2 achieves a consistently higher $SepSc$ for all these pairs. This demonstrates that our method effectively enhances the model's ability to discriminate between the confusable reasoning structures. See further analyses in Appendix E.

5.5 Enhancement of Deterministic Reasoning

To quantitatively assess the improvement in reasoning determinism, we construct a challenge test set comprising questions that are prone to cause confusion. We analyze the model behaviors on this set by performing temperature sampling for both MemQ and DR^2. A model is considered to perform deterministic reasoning if it generates a path with a probability greater than 90%.

The histogram in Figure 4 presents the distribution of probabilities that the model generates the correct reasoning path. The result shows that DR^2 increases the probability of deterministically inferring the correct path by 65% over MemQ. This confirms that our method effectively enables the model to perform deterministic reasoning towards the correct path.

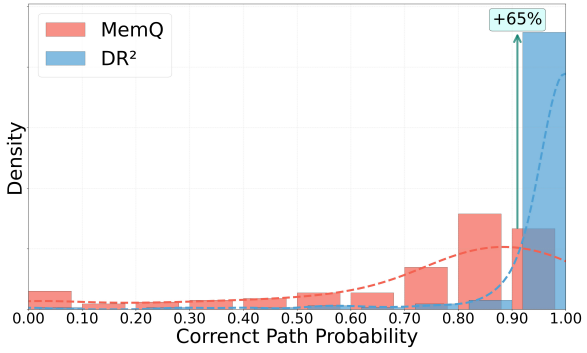

Figure 4: Comparing reasoning determinism of MemQ and DR^2 via temperature sampling on a confusion set.

5.6 Cross-Model Effectiveness

To assess the generalizability of DR^2 beyond a single model, we evaluate its performance when using different base models. We conduct experiments on four widely-used LLMs, the results in Table 4 demonstrate that all evaluated models achieve strong performance. This indicates that our approach to diagnosing and remedying semantic representation deficiencies is not restricted to a single model architecture, but instead exhibits effective transferability across a variety of backbone models. These findings further establish DR^2 as a robust and versatile framework for enhancing deterministic reasoning in KGQA, regardless of the underlying language model.

5.7 Ablation Study

To validate the contribution of each component in our proposed DR^2 method, we conduct a

Base	Method	WebQSP			CWQ		
		Hits@1	F1	Hits	Hits@1	F1	Hits
Llama2	MemQ	0.847	0.860	0.890	0.805	0.836	0.878
	DR^2	0.882 (+0.035)	0.889 (+0.029)	0.918 (+0.028)	0.835 (+0.030)	0.860 (+0.024)	0.898 (+0.020)
Llama3	MemQ	0.868	0.874	0.895	0.814	0.839	0.879
	DR^2	0.882 (+0.014)	0.888 (+0.014)	0.919 (+0.024)	0.840 (+0.026)	0.868 (+0.029)	0.911 (+0.032)
Qwen	MemQ	0.830	0.845	0.877	0.796	0.822	0.868
	DR^2	0.840 (+0.010)	0.857 (+0.012)	0.899 (+0.022)	0.808 (+0.012)	0.837 (+0.015)	0.889 (+0.021)
Gemma	MemQ	0.854	0.856	0.879	0.799	0.826	0.867
	DR^2	0.870 (+0.016)	0.876 (+0.020)	0.905 (+0.026)	0.809 (+0.010)	0.839 (+0.013)	0.883 (+0.016)

Table 4: Evaluation of the generalization of DR^2 across multiple widely-used LLMs.

comprehensive ablation study with the following four experimental settings: **1) Without Inverse Question Generation (w/o IQG).** During the preference construction stage, only the forward preference data $D_{forward}$ are used for DPO training, thereby isolating the contribution of the synthesized counterfactual data. **2) Without Semantic Distinguishability Evaluation (w/o SDE).** It eliminates the semantic distinguishability evaluation in the diagnosis stage. All summarized structural confusion pairs are directly used as the target defect set T_{Defect} for subsequent correction. **3) Without Confusion Error Summarization (w/o CES).** It uses all sampled confusion error instances from D_{Error} are directly used to construct forward preferences for DPO training. **4) Without Direct Preference Optimization (w/o DPO).** The model is directly fine-tuned on the original training set D_{Train} using a SFT objective.

Strategy	WebQSP			CWQ		
	Hits@1	F1	Hits	Hits@1	F1	Hits
DR^2	**0.882**	**0.889**	**0.918**	**0.835**	**0.860**	**0.898**
w/o IQG	0.875	0.882	0.913	0.820	0.850	0.892
w/o SDE	0.864	0.873	0.905	0.819	0.850	0.889
w/o CES	0.844	0.861	0.904	0.812	0.845	0.887
w/o DPO	0.848	0.861	0.891	0.801	0.833	0.876

Table 5: Ablation study on the components of DR^2.

Based on the results in Table 5, we observe the following: 1) The performance gap between DR^2 and "w/o IQG" demonstrates that the generated inverse preference data are essential for capturing fine-grained semantic distinctions. 2) The further decline in performance from "w/o IQG" to "w/o SDE" highlights the necessity of semantic distinguishability evaluation; without this step, the refinement process is adversely affected by confusion pairs unrelated to inherent model deficiencies. 3) The comparison between "w/o Semantic Evaluation" and "w/o Confusion Summarization" reveals that using all error instances without abstraction impedes the model's ability to focus on the most frequent and systematic confusion pairs, resulting in diminished performance. 4) Most notably, the largest performance drop occurs in the "w/o DPO" setting, confirming that standard SFT alone is inadequate for addressing the underlying semantic representation deficiencies in LLMs.

5.8 Reasoning Precision Improvement

Total Hops	1	2	3,4	>=5	avg
EHR					
KaeDe	0.716	0.770	0.759	0.679	0.741
MemQ	0.800	0.817	0.828	0.839	0.821
DR^2	**0.875**	**0.867**	**0.866**	**0.869**	**0.868**
GoldGED					
KaeDe	0.157	0.420	0.744	2.078	0.654
MemQ	0.170	0.358	0.659	1.118	0.523
DR^2	**0.155**	**0.315**	**0.619**	**1.075**	**0.490**

Table 6: Evaluation of edge accuracy (EHR) and search structure ($GoldGED$) for DR^2 and baselines.

To further investigate the improvement in the model's reasoning capability, we evaluate the reasoning paths generated by the model against the golden reasoning paths. The evaluation focuses on two key aspects: 1) the accuracy of the edges and 2) the accuracy of the search structure.

We use the Edge Hit Rate (EHR) to evaluate the accuracy of edges, which is defined as the proportion of edges in the golden path that are correctly predicted in the predicted path.

$$EHR = \frac{num(\{r|r \in P_{pred} \wedge r \in P_{gold}\})}{num(\{r|r \in P_{gold}\})}. \quad (6)$$

We use the Graph Edit Distance ($GoldGED$) to evaluate the accuracy of the search structure, which is defined as the edit distance between the predicted

path and the golden path. A lower $GoldGED$ indicates a higher structural accuracy.

$$GoldGED = \min_{\pi \in \Pi(P_{pred}, P_{gold})} num(\pi). \quad (7)$$

The evaluation results are presented in Table 6. In the comparison of edge accuracy, DR^2 method achieves a significantly higher EHR than all baseline methods. Moreover, as the complexity of the questions increases, the EHR of our method remains at a comparable level without a noticeable decline, demonstrating its robustness. In the comparison of graph structure, our method attains a lower $GoldGED$ than baselines, confirming that the search graphs generated by our approach predicts reasoning path structures more accurately.

5.9 Error Analysis

To gain deeper insight into the specific improvements introduced by our method, we conduct a detailed error analysis on both datasets. We categorize errors according to the source of reasoning confusion: **(1) Main Path Confusion**, where errors occur along the primary reasoning path from the starting entity to the answer entity; and **(2) Filter Path Confusion**, which arises from the misapplication of filtering constraints.

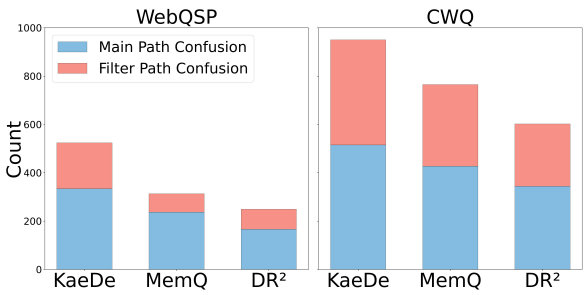

Figure 5: Error analysis of main path and filter path confusion for DR^2 and baselines.

As shown in Figure 5, DR^2 yields fewer errors in both Main Path and Filter Path confusion compared to the strong baselines MemQ and KaeDe. This demonstrates that the enhanced fine-grained semantic discrimination enabled by our method effectively reduces core relation misidentification and improves the precision of constraint application, thereby leading to more robust and reliable reasoning.

5.10 Case study

We present a case to intuitively demonstrate how DR^2 enhances deterministic reasoning. As shown in Figure 6, for the query *"Where did Thomas Hobbes live?"*, the golden reasoning path is *"Find the location where Thomas Hobbes has lived"*. The structural pair *"the location where someone has lived"* and *"the place of birth of someone"* was identified as a semantic representation deficiency during diagnosis. In this case, DR^2 demonstrates clear deterministic reasoning by assigning a high probability of 97.56% to the correct path, while MemQ exhibits pronounced indecision, assigning 51.92% to the incorrect path *"Find the place of birth of the person Thomas Hobbes"* and 44.53% to the correct one. See more cases in Appendix F.

WebQTest-1370

Question: where did thomas hobbes live?
Golden Path:
Find where *Thomas Hobbes* has lived.

DR^2

Generated Path: ($P = 97.56\%$)
Find where *Thomas Hobbes* has lived.

MemQ

Generated Path 1: ($P = 51.92\%$)
Find place of birth of *Thomas Hobbes*.
Generated Path 2: ($P = 44.53\%$)
Find where *Thomas Hobbes* has lived.

Figure 6: Case study comparing the deterministic reasoning of DR^2 and MemQ on a representative query.

6 Conclusion

In this paper, we propose DR^2, a novel method that mitigates the key challenge of non-deterministic reasoning in KGQA by diagnosing and remedying underlying representation deficiencies. Our approach diagnoses such deficiencies by analyzing the model's non-deterministic behavior under temperature sampling. These identified deficiencies are remedied by constructing targeted preference data through abductive reasoning-based Preference Learning. Experimental results demonstrate that DR^2 has achieved the state-of-the-art performance on WebQSP and CWQ, confirming the effectiveness of diagnosing and remedying representation deficiencies. These results not only demonstrate the practical advantages of our method, but also highlight the fundamental importance of semantic representation robustness in complex reasoning tasks with LLMs.

Limitation

Although our proposed DR^2 method demonstrates strong performance in the KGQA task by mitigating the issue of non-deterministic reasoning, we acknowledge several limitations in the present work that point to directions for future improvement:

1) Dependency on Labeled Data: While our proposed DR^2 method effectively diagnoses representation deficiencies, the process requires a substantial amount of labeled data. Specifically, identifying deficiencies requires sampling multiple reasoning paths for each query and comparing them with the annotated path. This reliance on annotated data to locate semantic ambiguities limits the method's scalability in low-resource scenarios. In future work, we will explore pathways to detect representation deficiencies directly from all relations in the knowledge graph, thereby reducing the reliance on labeled data and enabling a more automated diagnosis.

2) Limited Validation on Task Generalization: Our current work evaluates the proposed diagnostic and remediation method primarily within the KGQA task, which validates its capability to resolve semantic ambiguities in this setting. However, its effectiveness on other tasks where LLMs exhibit similar representation deficiencies remains unexplored. For instance, in tasks like Code Generation, Relation Extraction and Sentiment Analysis, models may also struggle with semantically similar options, leading to non-deterministic reasoning. It remains an open question whether the proposed deficiency diagnosis and remediation method can transfer to these tasks. In future work, we will explore the adaptability of our approach across a wider range of tasks.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Ranran Bu, Jian Cao, Jianqi Gao, Shiyou Qian, and Hongming Cai. 2025. KaeDe: Progressive generation of logical forms via knowledge-aware question decomposition for improved KBQA. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10958–10973, Suzhou, China. Association for Computational Linguistics.

Jianqi Gao, Jian Cao, Ranran Bu, Nengjun Zhu, Wei Guan, and Hang Yu. 2025. Promoting knowledge base question answering by directing llms to generate task-relevant logical forms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23914–23922.

Yu Gu, Xiang Deng, and Yu Su. 2023. Don't generate, discriminate: A proposal for grounding language models to real-world environments. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 4928–4949.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Huang, Sitao Cheng, Shanshan Huang, Jiayu Shen, Yong Xu, Chaoyun Zhang, and Yuzhong Qu. 2024. QueryAgent: A reliable and efficient reasoning framework with environmental feedback based self-correction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5014–5035.

Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, Yang Song, Chen Zhu, Hengshu Zhu, and Ji-Rong Wen. 2025. KG-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9505–9523, Vienna, Austria. Association for Computational Linguistics.

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

Yunshi Lan and Jing Jiang. 2020. Query graph generation for answering multi-hop complex questions from knowledge bases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 969–974, Online. Association for Computational Linguistics.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023. Few-shot in-context learning on knowledge base question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6966–6980.

Xiao Long, Liansheng Zhuang, Aodi Li, Minghong Yao, and Shafei Wang. 2025. Eperm: An evidence path enhanced reasoning model for knowledge graph question and answering. In *Proceedings of the AAAI*

Conference on Artificial Intelligence, volume 39, pages 12282–12290.

Haoran Luo, E Haihong, Zichen Tang, Shiyao Peng, Yikai Guo, Wentai Zhang, Chenghao Ma, Guanting Dong, Meina Song, Wei Lin, et al. 2024. Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2039–2056.

LINHAO LUO, Yuan-Fang Li, Reza Haf, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

Zhijie Nie, Richong Zhang, Zhongyuan Wang, and Xudong Liu. 2024. Code-style in-context learning for knowledge-based question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18833–18841.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Yuan Sui, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi. 2025. FiDeLiS: Faithful reasoning in large language models for knowledge graph question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8315–8330, Vienna, Austria. Association for Computational Linguistics.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4231–4242.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651.

Xiaqiang Tang, Jian Li, Nan Du, and Sihong Xie. 2025. Adapting to non-stationary environments: Multi-armed bandit enhanced retrieval-augmented generation on knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12658–12666.

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

Derong Xu, Xinhang Li, Ziheng Zhang, Zhenxi Lin, Zhihong Zhu, Zhi Zheng, Xian Wu, Xiangyu Zhao, Tong Xu, and Enhong Chen. 2025a. Harnessing large language models for knowledge graph question answering via adaptive multi-aspect retrieval-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25570–25578.

Mufan Xu, Gewen Liang, Kehai Chen, Wei Wang, Xun Zhou, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025b. Memory-augmented query reconstruction for LLM-based knowledge graph reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24068–24084, Vienna, Austria. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. 2022. RNG-KBQA: Generation augmented iterative ranking for knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6032–6043, Dublin, Ireland. Association for Computational Linguistics.

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.

Donghan Yu, Sheng Zhang, Patrick Ng, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Yiqun Hu, William Yang Wang, Zhiguo Wang, and Bing Xiang. 2022. Decaf: Joint decoding of answers and logical forms for question answering over knowledge bases. In *The Eleventh International Conference on Learning Representations*.

Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. 2022. Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5773–5784.

Xinghua Zhang, Haiyang Yu, Cheng Fu, Fei Huang, and Yongbin Li. 2025. IOPO: Empowering LLMs with complex instruction following via input-output preference optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22185–22200, Vienna, Austria. Association for Computational Linguistics.

A Error Analysis of MemQ

In order to understand the limitations of the strong baseline method MemQ, we conducted a manual error analysis on its incorrect predictions within the first 500 samples of WebQSP dataset. The observed errors are systematically categorized into four primary types: **1) Relation Confusion.** It predicts a knowledge graph relation that is semantically related to the correct one; **2) Rare Relation.** The target relation is highly sparse or unseen in the training data; **3) Constraint Violation.** It incorrectly filters out valid answers by over-applying constraints; **4) Query Ambiguity.** The input natural language query is ambiguous or underspecified. As illustrated in Figure 7, Relation Confusion is the most prevalent failure, accounting for 53% of the analyzed errors. This suggests that the model's primary weakness lies in distinguishing between semantically similar relations.

B Limitation of SFT

To empirically investigate the limitations of supervised fine-tuning (SFT) in discriminating closely related relations, we design a controlled experiment. We construct a dedicated dataset

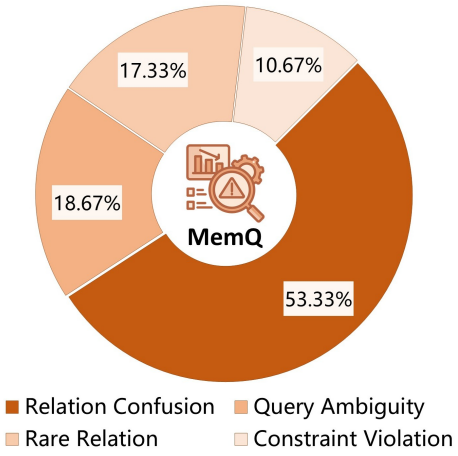

Figure 7: Error Analysis of MemQ.

where, for each query, the golden reasoning path must be exactly one of the two semantically similar relations: "language spoken" or "official language". This setup isolates the model's ability to discriminate between easily confusable relations without interference from other error types.

We randomly split this dedicated dataset into training and test sets, with the results shown in Figure 8. The performance is measured by the average probability it assigns to generating the correct golden reasoning path. Although the SFT model achieves near-perfect accuracy on the training set, its performance drops to 75% on the test set. This discrepancy indicates that the SFT objective leads the model to primarily memorize surface patterns rather than learn to discriminate between closely related relations.

C Evaluation Metrics

In this section, we present the mathematical formulations and explanations for the primary metrics used in our evaluation. All reported results are averaged values.

Hits@1. Hits@1 quantifies the proportion of questions for which the top-ranked answer in the model's output is correct. Let $Answer$ represent the list of predicted answers, $Golden$ denote the list of ground truth answers, and $total$ represent the total number of questions in the dataset. The formula of Hits@1 is defined as follows:

$$Hits@1 = \frac{count(Answer[0] \in Golden)}{total}. \quad (8)$$

Hits. The Hits metric measures whether **any** of the model's predicted answers is present in the set of

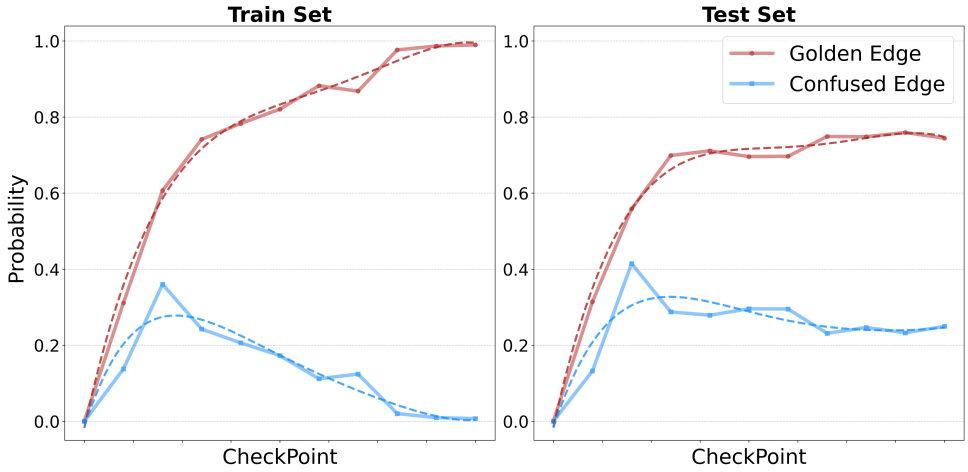

Figure 8: Limitation of SFT.

ground truth answers. It is formulated as:

$$Hits = \frac{count(Answer \cap Golden \neq \emptyset)}{total}. \quad (9)$$

Unlike Hits@1, which is a stricter measure requiring the first prediction to be correct, Hits considers the prediction successful if any of the returned answers matches a golden answer.

F1. The F1 score provides a balanced measure of the model's overall answer quality by considering both precision and recall. We report the overall F1 score. This is calculated by first averaging the precision and recall values across all samples and then computing the F1 score based on them.

D Prompt Templates

The prompt templates used in our experiments are detailed below. We design two specific templates: one for the Semantic Distinguishability Evaluation (see Table 8) and another for the Inverse Question Generation via abductive reasoning (see Table 9).

E Further Representation Analysis

We provide additional evidence to illustrate how DR^2 mitigates representation deficiencies. This appendix presents extended analyses and visualizations on the identified defect set T_{Defect}, comparing the internal representations of DR^2 against those of the MemQ baseline. Figure 11 further demonstrate DR^2's effectiveness in separating previously confused relations.

Compared to MemQ, DR^2 consistently produces more distinct and better-separated representation clusters for the two reasoning structures within each confusable pair, with a clearer boundary between them. This clear separation observed in the visualizations is supported by the quantitative results. For every defect pair examined, DR^2 achieves a significantly higher Separability Score ($SepSc$) than MemQ. This combined evidence robustly demonstrates that our method effectively enhances the model's ability to discriminate between previously confusable reasoning structures.

F Extended Case Studies

To further illustrate how DR^2 enhances deterministic reasoning, we present the following extended case studies.

In Figure 9, the query is *"what country does rafael nadal play for?"*. The baseline SFT method (MemQ) exhibits confusion among multiple semantically related knowledge graph relations: *"nationality of the person someone"*, *"the country of which someone is a notable person"*, and *"the country of the country where someone has lived"*. exhibiting non-deterministic reasoning. In contrast, DR^2 maintains deterministic and correct reasoning reasoning, successfully selecting the precise relation *"nationality of the person someone"*. This case highlights DR^2's enhanced ability to resolve multi-relation confusion through improved semantic discrimination.

In Figure 10, the query is *"what region of the world is egypt associated with?"*. Although the baseline MemQ correctly predicts the relation *"the location that contains somewhere"*, it fails to apply the necessary constraint filtering for the *"notable type"*. In contrast, DR^2 accurately perceives the contextual cue *"region"* in the query and accordingly applies the necessary *"notable type"*

constraint during reasoning. This demonstrates DR^2's enhanced ability to handle constraint-aware reasoning, effectively resolving relation confusion that arises within filtering constraints.

G Details of Datasets

In this section, we provide details about the datasets utilized in this paper in Table 7. Both datasets are publicly available and pose no security or privacy concerns.

Dataset	Train	Test
WebQSP	3098	1639
CWQ	27639	3531

Table 7: Details about datasets.

H Use of AI in Writing

During the writing process, AI assistants were exclusively used for text editing purposes, including grammar correction and wording refinement.

WebQTest-1770

Question: what country does rafael nadal play for?
Golden Reasoning Path:
Step1: Find the nationality of the person *Rafael Nadal*, assign it to ?x.
Step2: Make sure ?x should not be *Rafael Nadal*.
Finally the answer is ?x.

DR^2

Generated Reasoning Path: ($P = 99.56\%$)
Step1: Find the nationality of the person *Rafael Nadal*, assign it to ?x.
Step2: Make sure ?x should not be *Rafael Nadal*.
Finally the answer is ?x.

MemQ

Generated Reasoning Path 1: ($P = 53.17\%$)
Step1: Find the nationality of the person *Rafael Nadal*, assign it to ?x.
Step2: Make sure ?x should not be *Rafael Nadal*.
Finally the answer is ?x.
Generated Reasoning Path 2: ($P = 27.90\%$)
Step1: Find the country of which *Rafael Nadal* is a notable person, assign it to ?x.
Step2: Make sure ?x should not be *Rafael Nadal*.
Finally the answer is ?x.
Generated Reasoning Path 3: ($P = 8.31\%$)
Step1: Find the country of the country where *Rafael Nadal* has lived, assign it to ?x.
Step2: Make sure ?x should not be *Rafael Nadal*.
Finally the answer is ?x.

Figure 9: Compare a case of DR^2 and MemQ.

WebQTest-376

Question: what region of the world is egypt associated with?
Golden Reasoning Path:
Step1: Find the location that contains *Egypt*, assign it to ?x.
Step2: Make sure ?x should not be *Egypt*.
Step3: Find the notable type of the topic ?x, assign it to ?notable_types.
Step4: Make sure ?notable_types should be *Region*.
Finally the answer is ?x.

DR^2

Generated Reasoning Path: ($P = 95.66\%$)
Step1: Find the location that contains *Egypt*, assign it to ?x.
Step2: Make sure ?x should not be *Egypt*.
Step3: Find the notable type of the topic ?x, assign it to ?notable_types.
Step4: Make sure ?notable_types should be *Region*.
Finally the answer is ?x.

MemQ

Generated Reasoning Path 1: ($P = 79.83\%$)
Step1: Find the location that contains *Egypt*, assign it to ?x.
Step2: Make sure ?x should not be *Egypt*.
Finally the answer is ?x.
Generated Reasoning Path 2: ($P = 13.18\%$)
Step1: Find the location that contains *Egypt*, assign it to ?x.
Step2: Make sure ?x should not be *Egypt*.
Step3: Find the notable type of the topic ?x, assign it to ?notable_types.
Step4: Make sure ?notable_types should be *Region*.
Finally the answer is ?x.

Figure 10: Compare a case of DR^2 and MemQ.

Semantic Distinguishability Evaluation

You are given two groups of questions:
Group1:
Grou1 Hints: {group1_structure}
Group1 Questions:
{group1_question1}
{group1_question2}
{group1_question3}
{group1_question4}
Group2:
Group2 Hints: {group2_structure}
Group2 Questions:
{group2_question1}
{group2_question2}
{group2_question3}
{group2_question4}
Here is another question, your task is to classify it into either Group1 or Group2.
Respond strictly with either "Group1" or "Group2". Do not provide any explanations or other text.
Question:
question

Table 8: Semantic Distinguishability Evaluation

Inverse Question Generation

Based on the examples below, generate ONLY the natural language question based on the provided Entity and its Search Plan. The question should accurately reflect the search intent, mimicking the style and phrasing of the examples provided below.
Output NOTHING BUT the Question itself.
Here are the examples:
EXAMPLE1
Entity: {topic_entity1}
Search Plan: {golden_plan1}
Question: {question1}
EXAMPLE2
Entity: {topic_entity2}
Search Plan: {golden_plan2}
Question: {question2}
EXAMPLE3
Entity: {topic_entity3}
Search Plan: {golden_plan3}
Question: {question3}
Your Task:
Entity: {topic_entity}
Search Plan: {generated_plan}
Question:

Table 9: Inverse Question Generation

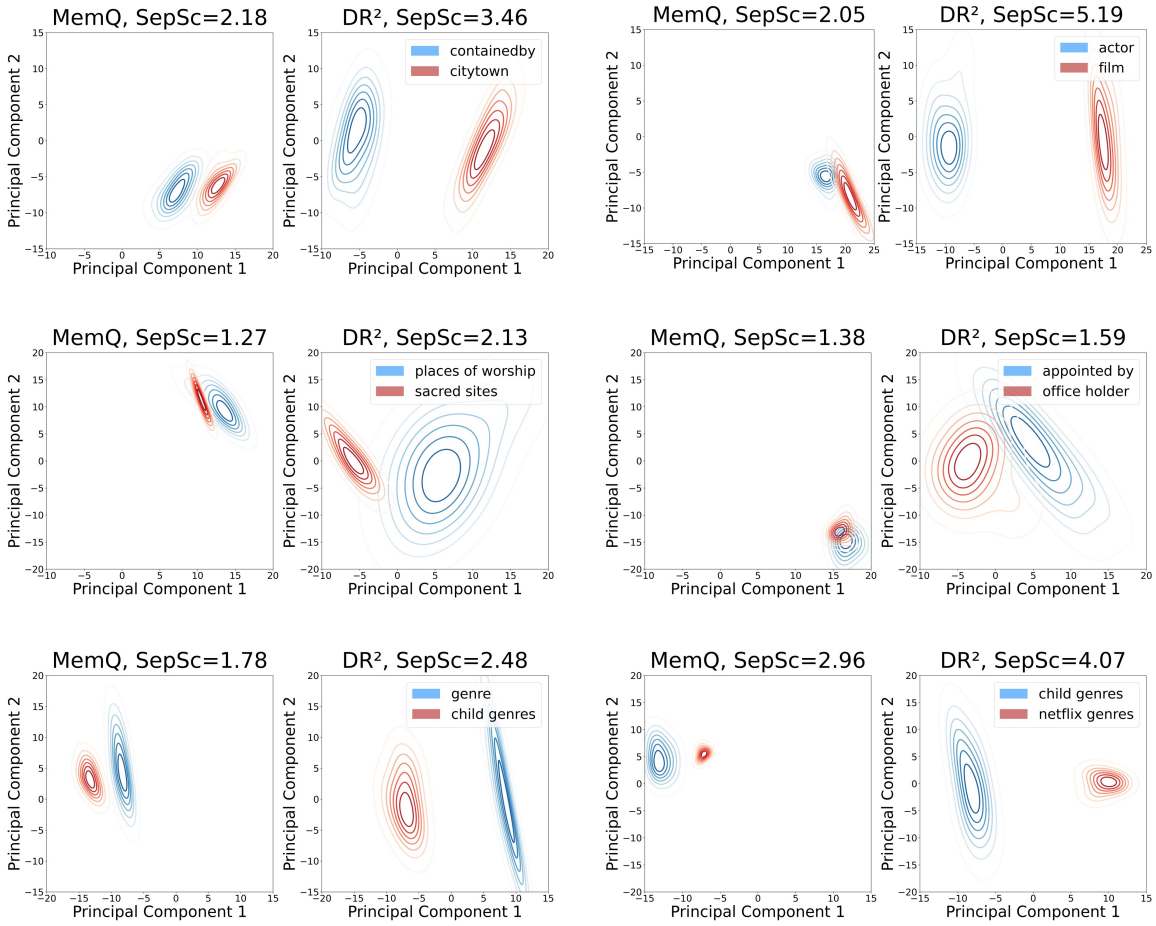

Figure 11: Further Representation Analysis.