What External Knowledge is Preferred by LLMs? Characterizing and Exploring Chain of Evidence in Imperfect Context for Multi-Hop QA

Anonymous ACL submission

Abstract

Incorporating external knowledge has emerged as a promising way to mitigate outdated knowledge and hallucinations in LLM. However, external knowledge is often imperfect, encompassing substantial extraneous or even inaccurate content, which interferes with the LLM's utilization of useful knowledge in the context. This paper seeks to characterize the features of preferred external knowledge and perform empirical studies in imperfect contexts. Inspired by the chain of evidence (CoE), we characterize that the knowledge preferred by LLMs should maintain both relevance to the question and mutual support among the textual pieces. Accordingly, we propose a CoE discrimination approach and conduct a comparative analysis between CoE and Non-CoE samples across significance, deceptiveness, and robustness, revealing the LLM's preference for external knowledge that aligns with CoE features. Furthermore, we selected three representative tasks (RAG-based multi-hop QA, external knowledge poisoning and poisoning defense), along with corresponding SOTA or prevalent baselines. By integrating CoE features, the variants achieved significant improvements over the original baselines.

1 Introduction

The parameterized knowledge acquired by large language models (LLMs) through pre-training at a specific point in time becomes outdated with the knowledge evolution or produces hallucination (Achiam et al., 2023; Touvron et al., 2023a; Anil et al., 2023). Incorporating external knowledge into LLM has emerged as an effective approach to mitigate this problem (Tu et al., 2024; Zhao et al., 2024). In this context, properties such as the accuracy and reliability of external knowledge are critical for LLMs to provide accurate answers.

However, external knowledge is often imperfect. In addition to the useful knowledge that users expect LLMs to follow, the context typically contains two types of noise (Chen et al., 2024; Zou et al., 2024): 1) extraneous information, despite showing textual similarities with the question, cannot support the correct answer (Chen et al., 2024; Xiang et al., 2024); 2) inaccurate information, which can mislead LLM to produce incorrect answers (Liu et al., 2024). Especially when dealing with complex scenarios such as multi-hop QA, the acquisition of such noise is inevitable due to limitations of retrievers or quality deficiencies in the specialized knowledge bases (Wang et al., 2024; Dai et al., 2024; Tang and Yang, 2024). This hinders LLMs from effectively using useful knowledge within external contexts and leads to incorrect answers.

Consequently, numerous studies aim to characterize the features of external knowledge that LLMs tend to follow in imperfect contexts (such as confirmation bias, completeness bias, coherent bias, etc.) (Xie et al., 2023; Zhang et al., 2024); or on approaches such as reranking or retrieval to prioritize knowledge with high relevance (Asai et al., 2023; Dong et al., 2024). However, previous studies primarily suffer from two main deficiencies. First, while their focus is on qualitative findings, it remains uncertain whether such findings can effectively guide performance improvements in representative tasks (Zhang et al., 2024). Second, their research focuses on single-hop QA, in which a single piece of knowledge suffices to answer the question. However, the generalizability of these findings to more complex scenarios (e.g., multi-hop QA) has yet to be confirmed.

In our study, we focus on characterizing what external knowledge is more capable of resisting the surrounding noise and guiding LLMs for better generation. Inspired by the Chain of Evidence (CoE) theory in criminal procedural law (Murphy, 2013), which requires case-decisive evidence to demonstrate both relevance (pertaining to the case) and interconnectivity (evidence mutually supporting each other) in judicial decisions. In multi-hop

QA, analogously to the scenario where LLMs rely on external knowledge for answering, we consider that the preferred knowledge should show relevance to the question (relevance) and mutual support and complementarity among textual pieces in addressing the question (interconnectivity). Based on the principle, we first characterize what knowledge can be considered CoE and propose a discrimination approach to determine whether the given external knowledge aligns to the CoE features. Subsequently, we conduct a comparative analysis of CoE versus Non-CoE samples, examining LLMs' preference for CoE-aligned content across four dimensions below.

- **Significance**, we examine whether LLMs demonstrate superior performance when provided with external knowledge exhibiting CoE characteristics, versus cases where the knowledge is relevant but lacks COE characteristics.
- **Deceptiveness**, we investigate whether samples that conform to COE characteristics but lead to incorrect answers exhibit higher deceptive potential, effectively inducing LLMs to generate incorrect output.
- **Robustness**, we investigate whether the approach effectively mitigates knowledge conflicts and enhances question-answering performance in multi-hop scenarios.
- **Usability**, we select three representative tasks (RAG-based multi-hop QA, external knowledge poisoning and defenses) to explore whether CoE can be effectively integrated and enhance the effectiveness of baselines.

Using HotpotQA (Yang et al., 2018) and 2WikiMultihopQA (Ho et al., 2020) as data sources, we constructed 1,336 multi-hop QA pairs and the corresponding CoE based on the automatic discrimination. By applying perturbations to CoE, we also build Non-CoE samples (that is, knowledge lacking the necessary relevance or interconnectivity to establish CoE) for each QA pair. Subsequently, we conducted a comprehensive evaluation in five state-of-the-art LLMs (GPT-3.5 (OpenAI, 2022), GPT-4 (Achiam et al., 2023), LLama2-13B (Touvron et al., 2023b), LLama3-70B (Touvron et al., 2023a), and Qwen2.5-32B (Qwen Team, 2024).

The empirical analysis implies that if external knowledge in the context exhibits CoE characterization, it can better resist interference from extraneous and even inaccurate information and improve multi-hop QA performance. Building upon these findings, we can effectively enhance existing multi-hop QA approaches and poisoning defenses through performance improvements. Nevertheless, the observed preference for CoE-compliant external knowledge creates a vulnerability. Adversaries can deliberately construct false information with CoE characteristics to successfully trick LLMs into generating answers containing factual errors. Empirically, our investigation uncovers characteristic preferences of LLMs toward external knowledge from both relevance and interconnectivity perspectives, which informs the optimization of knowledge representation and retrieval mechanisms in RAG systems. Practically, our studies demonstrate significant improvements over the SOTA or prevalent baselines in three representative tasks. The reproduction package is available at: `https://anonymous.4open.science/r/ScopeCOE-78D3`.

2 Related Work

In imperfect knowledge augmentation, there is growing interest in understanding LLMs' knowledge preferences, especially in contexts involving conflicts between external and internal knowledge, as well as contradictions within internal knowledge (Xie et al., 2023; Kasai et al., 2023; Tan et al., 2024; Jin et al., 2024; Xu et al., 2024b,a).

Xie et al. (2023) demonstrated LLMs' bias towards coherent knowledge, revealing that LLMs are highly receptive to external knowledge when presented coherently, even when it conflicts with their parametric knowledge. Jin et al. (2024) found that LLMs demonstrate confirmation bias, manifested as their inclination to choose knowledge consistent with their internal memory, regardless of whether it is correct or incorrect. Chen et al. (2022) demonstrated LLMs' preference for highly relevant knowledge by manipulating retrieved snippets based on attention scores, showing that LLMs prioritize knowledge with greater relevance to questions. Zhang et al. (2024) found LLMs perform better when given complete external knowledge, showing completeness bias.

Although existing studies have documented LLMs' knowledge preferences, there exists a significant gap in understanding and measuring the essential features that govern these preferences, especially in complex scenarios like multi-hop QA.

To this end, we manage to characterize and discriminate external knowledge that can help LLMs generate correct responses.

3 CoE Characterization and Discrimination

3.1 CoE Characterization

Drawing from the law of criminal procedure, judicial decisions in cases require the formation of a CoE through evidence collection (Edmond and Roach, 2011; Murphy, 2013). Such a CoE must demonstrate two properties: relevance (pertaining to the case) and interconnectivity (evidence mutually supporting each other). In multi-hop QA, the user question is analogous to a legal case, where external contexts constitute the evidentiary collection, and the LLM's answer represents the judicial conclusion drawn through iterative reasoning processes. Based on this analogy, we hypothesize that in the reasoning process from user query to final answer, LLMs tend to prioritize external knowledge that demonstrates both relevance and interconnectivity.

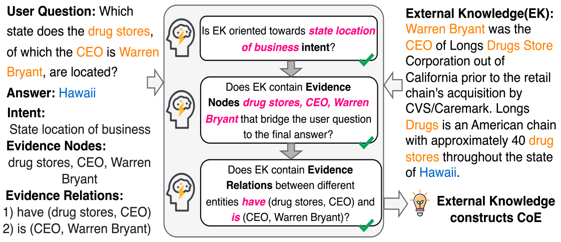

User Question: Which state does the drug stores, of which the CEO is Warren Bryant, are located?
Answer: Hawaii
Intent:
State location of business
Evidence Nodes:
drug stores, CEO, Warren Bryant
Evidence Relations:
1) have (drug stores, CEO)
2) is (CEO, Warren Bryant)
Is EK oriented towards state location of business intent?
Does EK contain Evidence Nodes drug stores, CEO, Warren Bryant that bridge the user question to the final answer?
Does EK contain Evidence Relations between different entities have (drug stores, CEO) and is (CEO, Warren Bryant)?
External Knowledge(EK): Warren Bryant was the CEO of Longs Drugs Store Corporation out of California prior to the retail chain's acquisition by CVS/Caremark. Longs Drugs is an American chain with approximately 40 drug stores throughout the state of Hawaii.
External Knowledge constructs CoE

Figure 1: Example of CoE and the CoE features.

Next, we will characterize relevance and interconnectivity from the following three perspectives of textual features presented by external knowledge.

- **Intent** is a noun or noun phrase representing the user's desired answer to their question, and it aims to align the purpose of the user's question with the ultimate facts derived from external knowledge.

- **Evidence Nodes** are the key entities in a user's question, which imply critical knowledge elements for multi-hop reasoning. It ensures logical consistency between the starting and ending points of a single reasoning hop, aligning the user's query with external knowledge.

- **Evidence Relations** are logical predicates within the question, indicating the semantic associations between each pair of evidence nodes. It is used to verify whether the implicit semantic connections between entities in external knowledge are consistent with the inherent logic in the question.

Taking Figure 1 as an example, intent specifies "state location of business" as the goal, indicating that the user wants to find the state where the business operates. The evidence nodes, "drug stores", "CEO", and "Warren Bryant", serve as essential nodes for multi-hop reasoning. Evidence relations show how these entities are linked, with "have" connecting "drug stores" to "CEO", and "is" linking "CEO" to "Warren Bryant". The effectiveness of CoE stems from the synergistic interaction of these three features. The integration of all three features creates a comprehensive logic chain tailored to the specific question.

3.2 CoE Discrimination

Based on the characterized features, we design an approach to discriminate whether external knowledge exhibits CoE features. Generally, discrimination extracts three CoE features from the user query, i.e., intent, evidence node, and evidence relation. These features represent the underlying logic embedded within the user question and serve as objectives for external knowledge alignment. Subsequently, for a given piece of external knowledge, we verify whether it simultaneously satisfies all three features. The following introduces the details for the implementation of CoE discrimination.

First, for a user question, we extract its intent, the evidence nodes and the evidence relations using GPT-4o with a hand-crafted prompt. Second, with the extracted CoE features, we discriminate whether external knowledge exhibits them using GPT-4o. As for intent discrimination, we frame it as a textual entailment task, where external knowledge as a premise and intent as a hypothesis. We reason whether the hypothesis holds on the basis of the given premise using GPT-4o with prompting. To discriminate the nodes and relations, we uniformly treat this as a classification task about the "containment" logic. In implementation, we manually construct distinct prompts for each type of feature, instructing GPT-4o to classify whether an external knowledge contains the extracted nodes and relations. Finally, external knowledge is considered aligned to the user question only when all required CoE features are present. The discrimination prompts are detailed in Appendix I.

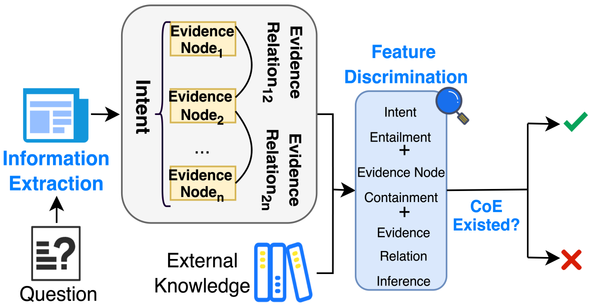


Figure 2: The overview of CoE discrimination.

4 Subject Dataset

4.1 CoE Knowledge Construction

Oriented to multi-hop QA, we selected two commonly used datasets, HotpotQA and 2WikiMultihopQA as sources. Both datasets contain samples comprising not just QA pairs, but also the supporting knowledge required to derive each answer. During construction, the supporting knowledge was specifically designed to capture the necessary logical chains for multi-hop reasoning. We consider it to be highly compatible with the COE features, thus qualifying it as a candidate COE knowledge.

Referring to the sample size in previous studies (Jin et al., 2024; Chen et al., 2024), we randomly sampled 1,000 instances from each dataset and applied the CoE discrimination approach to check whether each candidate exhibits CoE features. Finally, we obtained 676 and 660 knowledge pieces that contain CoE from candidates, with an average of 4.0 and 3.4 sentences within the supporting knowledge for two datasets, respectively.

4.2 Non-CoE Knowledge Construction

Based on the CoE Knowledge, we construct Non-CoE knowledge using two controlled perturbation strategies that are commonly employed in empirical studies of LLM knowledge construction (Xie et al., 2023): sentence-level perturbation (SenP) and word-level perturbation (WordP). SenP simulates incomplete knowledge by removing key evidence pieces, while WordP replaces specific evidence nodes with their higher-level expressions. These strategies ensure fair comparison by maintaining the same question context while only varying the CoE completeness. Detailed perturbation procedures and examples are provided in Appendix E. Subsequently, we construct complete LLM contexts by augmenting COE and Non-COE knowledge with extraneous and inaccurate information. Through comparative analysis, we systematically examine their performance differences across various scenarios.

5 Significance Assessment

In practice, the context of LLMs often contains non-trivial noise due to limitations in knowledge base quality or retriever performance (Liu et al., 2024). This results in situations where even when useful knowledge is retrieved, other noise may interfere with it, ultimately preventing LLMs from generating desired content. In this section, we aim to investigate whether knowledge that aligns with CoE features can more effectively resist noise in the context and help LLMs answer multi-hop questions. To this end, we employ a comparative analysis methodology. We retrieve text fragments that are lexically similar but offer limited relevance to QA, simulating potential extraneous information interference in contexts. These fragments are then combined with either CoE or Non-CoE knowledge to form complete contexts, followed by an analysis of whether significant performance differences exist between the two groups.

5.1 Experimental Design

We design a comprehensive experimental framework to evaluate the effectiveness of CoE and Non-CoE (Non-CoE$_{SenP}$ and Non-CoE$_{WordP}$) under different noise conditions. Our analysis focuses on four key dimensions: (1) the performance comparison between CoE and Non-CoE on LLMs, (2) the impact of varying extraneous ratios on their effectiveness, and (3) the performance of CoE in single-hop and multi-hop scenarios. (4) the influence of different CoE features on LLM performance. We inject extraneous information at four different ratios (from 0 to 0.75, with 0.25 intervals) to examine how each approach maintains its effectiveness.

For evaluation, we select five representative LLMs that span both closed-source (GPT-3.5, GPT-4) and open-source LLMs (LLama2-13B, LLama3-70B, Qwen2.5-32B) to ensure comprehensive coverage across different model scales and architectures. Following general QA evaluation protocols Adlakha et al. (2024), we use GPT-4 as the judge to compute the accuracy (ACC) between model outputs and ground truth answers. To understand CoE's significance more comprehensively, we conduct

Table 1: LLMs' Accuracy (ACC) on CoE and Non-CoE.

Model	Irrelevant Proportion	HotpotQA			2WikiMultihopQA		
		CoE	Non-CoE		CoE	Non-CoE	
			SenP	*WordP*		*SenP*	*WordP*
GPT-3.5	*0*	**91.9%**	77.9%*	79.1%*	**97.4%**	74.1%*	83.5%*
	0.25	**90.3%**	75.6%*	77.5%*	**96.9%**	68.2%*	81.2%*
	0.5	**89.9%**	73.1%*	75.4%*	**96.5%**	66.4%*	82.6%*
	0.75	**88.9%**	65.7%*	74.5%*	**95.4%**	58.4%*	70.8%*
GPT-4	*0*	**93.5%**	83.4%*	86.4%*	**93.7%**	67.7%*	79.4%*
	0.25	**93.4%**	82.3%*	86.4%*	**94.0%**	70.9%*	80.1%*
	0.5	**91.8%**	82.0%*	86.5%*	**95.4%**	71.5%*	77.3%*
	0.75	**91.2%**	80.1%*	83.8%*	**95.9%**	64.9%*	74.4%*
Llama2-13B	*0*	**89.9%**	87.1%*	88.8%*	**96.5%**	95.3%*	93.3%*
	0.25	**87.9%**	84.2%*	85.2%*	**95.9%**	93.7%*	91.9%*
	0.5	**86.4%**	82.8%*	84.0%*	**93.8%**	91.2%*	90.0%*
	0.75	**85.8%**	79.5%*	82.9%*	**90.9%**	86.6%*	86.3%*
Llama3-70B	*0*	**92.5%**	76.8%*	74.5%*	**95.7%**	79.5%*	73.3%*
	0.25	**92.9%**	74.1%*	76.1%*	**93.7%**	80.3%*	71.4%*
	0.5	**91.1%**	72.6%*	76.8%*	**95.9%**	76.7%*	69.6%*
	0.75	**90.5%**	69.8%*	68.3%*	**93.1%**	72.3%*	67.3%*
Qwen2.5-32B	*0*	**87.8%**	71.3%*	75.7%*	**90.7%**	53.1%*	67.0%*
	0.25	**87.2%**	38.6%*	64.9%*	**91.3%**	29.5%*	49.4%*
	0.5	**86.1%**	37.7%*	64.3%*	**92.1%**	27.8%*	47.5%*
	0.75	**88.0%**	37.3%*	57.2%*	**91.9%**	22.2%*	45.9%*

* indicates statistical significance compared to CoE ($p < 0.05$)

experiments on a single-hop dataset and perform ablation studies by removing different CoE features. Detailed experimental settings are provided in Appendix B and C.

5.2 Results and Discussion

Table 1 shows the ACC in different ratios of extraneous information. Comparing the CoE and Non-CoE groups, the results show that CoE achieves an average ACC of 92.0% across five LLMs and two datasets, outperforming Non-CoE$_{SenP}$ and Non-CoE$_{WordP}$ by 22.5% and 16.3%, respectively. This substantial improvement suggests that external knowledge that exhibiting CoE features enables LLMs to utilize it and achieve better performance.

For the proportion of extraneous information in the context, as the ratio increases from 0% to 75%, CoE's ACC only decreases by 1.8%, while the ACC of Non-CoE variants decreases more significantly: 12.9% for Non-CoE$_{SenP}$ and 9.0% for Non-CoE$_{WordP}$. This robustness suggests that external knowledge that exhibits CoE features helps LLMs maintain consistent comprehension and reasoning faced with noise of different magnitudes.

Analyzing the impact of reasoning complexity, experiments show that single-hop reasoning maintains the most stable performance (>92.0% ACC) under increasing extraneous information, followed by three-hop reasoning (>90.0% ACC), while two-hop reasoning exhibits higher sensitivity, with ACC dropping from 91.0% to 88.0%. This pattern suggests that CoE is particularly effective in simpler reasoning scenarios, while maintaining strong performance in more complex cases. Detailed experimental results and analysis are provided in Appendix C.

In addition, ablation studies further demonstrate the varying impacts of CoE features: removing intent causes the largest accuracy drop (33.9%), followed by evidence nodes (13.6%), while evidence relation removal has the smallest impact (10.7%). It indicates that explicit reasoning intent is crucial for guiding LLMs' responses. Detailed experimental results and analysis are provided in Appendix B.

Summary: If external knowledge exhibits CoE characterization, it can better resist interference from extraneous information and improve multi-hop QA performance. Moreover, LLMs exhibit greater resistance if there exists external knowledge exhibiting CoE features, as the proportion of extraneous information increases. For practical guidance, optimizing the retriever to prioritize knowledge exhibiting CoE features can effectively enhance performance of multi-hop QA.

6 Deceptiveness Assessment

Given that CoE represents structured reasoning chains, it is crucial to examine whether such well-formed evidence paths could amplify the deceptive effect of poisoned knowledge. Therefore, we investigate a more challenging scenario, where the CoE contains factual errors, to determine whether LLMs can still be effectively misled and produce answers consistent with the incorrect information embedded in the CoE.

6.1 Experimental Design

To investigate the deceptiveness of incorrect external knowledge under imperfect conditions, we design a comprehensive evaluation framework comparing CoE and Non-CoE. Our analysis focuses on three key dimensions: (1) the comparative effectiveness between CoE and Non-CoE (Non-CoE$_{SenP}$ and Non-CoE$_{WordP}$) in misleading LLMs with incorrect information, and (2) how their deceptive capabilities change under varying ratios of irrelevant information, and (3) the influence of different CoE features on LLM deception effectiveness.

In constructing incorrect information for both CoE and Non-CoE, we carefully preserve the semantic type and format of original answers (e.g., replacing "United States" with "Canada" while maintaining consistent structures) to ensure fair

Table 2: Attack Success Rate (ASR) of CoE and Non-CoE against LLMs.

Model	Irrelevant Proportion	HotpotQA			2WikiMultihopQA		
		CoE	Non-CoE		CoE	Non-CoE	
			SenP	*WordP*		*SenP*	*WordP*
GPT-3.5	*0*	**86.1%**	75.6%*	83.1%*	**85.0%**	58.5%*	57.4%*
	0.25	**85.8%**	76.0%*	79.1%*	**86.5%**	53.8%*	52.4%*
	0.5	**84.7%**	72.2%*	77.8%*	**84.2%**	50.0%*	48.8%*
	0.75	**78.4%**	72.0%*	73.7%*	**83.3%**	45.2%*	44.9%*
GPT-4	*0*	**86.5%**	52.2%*	59.0%*	**85.4%**	68.8%*	76.2%*
	0.25	**85.5%**	50.5%*	58.9%*	**87.2%**	67.0%*	73.2%*
	0.5	**84.0%**	46.8%*	52.7%*	**90.6%**	65.2%*	76.8%*
	0.75	**78.2%**	43.2%*	50.5%*	**92.7%**	62.3%*	75.1%*
Llama2-13B	*0*	**78.2%**	76.9%*	72.9%*	**91.5%**	89.8%*	88.6%*
	0.25	**77.1%**	74.1%*	67.3%*	**89.8%**	87.5%*	86.3%*
	0.5	**71.6%**	70.0%*	67.5%*	**89.1%**	86.8%*	85.1%*
	0.75	**69.1%**	64.5%*	64.8%*	**84.1%**	81.6%*	82.1%*
Llama3-70B	*0*	**82.8%**	76.9%*	72.8%*	**89.7%**	77.1%*	72.1%*
	0.25	**81.6%**	75.1%*	71.9%*	**89.5%**	72.1%*	70.4%*
	0.5	**78.0%**	71.7%*	68.0%*	**88.9%**	69.4%*	66.5%*
	0.75	**78.2%**	62.9%*	64.1%*	**89.8%**	51.4%*	53.7%*
Qwen2.5-32B	*0*	**90.6%**	68.9%*	79.1%*	**93.7%**	43.5%*	65.8%*
	0.25	**87.7%**	67.3%*	80.0%*	**93.6%**	47.2%*	67.3%*
	0.5	**86.3%**	64.1%*	76.5%*	**93.1%**	47.0%*	68.6%*
	0.75	**85.8%**	62.9%*	74.2%*	**94.0%**	46.5%*	65.6%*

* indicates statistical significance compared to CoE ($p < 0.05$)

comparison. Following Section 5.1, we inject irrelevant information at the same ratios to examine how each approach maintains its deceptive effectiveness.

Using the same subject LLMs as Section 5.1 and following standard evaluation protocols Adlakha et al. (2024), we use GPT-4o as the judge to compute the Attack Success Rate (ASR), defined as the proportion of successfully misled LLM outputs. We also conduct ablation studies to analyze how different CoE features midleading LLM responses. Detailed experimental settings are provided in B.

6.2 Results and Discussion

Table 3 shows the ASR of LLMs with external knowledge under CoE and two types of Non-CoE samples leading to incorrect answers. The results show that the average ASR reaches 85.4% for the COE group, which is 20.6% and 16.2% higher than Non-CoE$_{SenP}$ and Non-CoE$_{WordP}$, respectively. The results imply that CoE demonstrates significant deception in misleading LLMs when it contains factual errors.

Combining with the ratio of the irrelevant information, as the ratio increases from 0% to 75%, CoE's ASR only decreases by 3.6%, while the attack effectiveness of Non-CoE variants drops more significantly (9.7% for Non-CoE$_{SenP}$ and 7.9% for Non-CoE$_{WordP}$). In general, CoE's attack effectiveness remains more stable against LLMs when faced with irrelevant noise variations, outperforming two types of Non-CoE samples. Another noteworthy finding is that when the CoE leads to an incorrect answer, its performance metrics are on average 6.6% lower than those of a correct CoE (Table 1). This phenomenon probably stems from the parametric knowledge inherent in LLMs, which confers a degree of resistance to poisoned or erroneous knowledge input.

Ablation studies reveal the varying impacts of CoE components on LLM deception: removing evidence nodes leads to the highest ASR (78.4%), followed by removing evidence relations (64.4%) and intent (53.6%). However, the absence of evidence nodes results in reduced stability against irrelevant knowledge, with ASR dropping by 9.4%, indicating their vital role in maintaining deceptive effectiveness under noisy scenarios. Detailed analysis is provided in Appendix B.

Summary: External knowledge exhibiting CoE characteristics demonstrates significant deceptiveness in misleading LLMs when it contains factual errors. This implies that external knowledge matching CoE features requires elevated prioritized safeguards due to their potent deceptiveness.

7 Robustness Assessment

In addition to examining CoE's performance when confronted with irrelevant information, we further investigate its resilience against knowledge conflicts in the context, e.g., cases where adversarial attacks have compromised other retrieved contexts.

7.1 Experimental Design

Starting from the CoE and Non-CoE samples, we continuously inject conflicting knowledge into their context at varying ratios to simulate scenarios where incorrect knowledge is retrieved or context is poisoned. We then observe and compare the performance of both sample groups in multi-hop QA to analyze whether CoE samples exhibit stronger robustness to misinformation interference.

First, we construct conflicting knowledge through two strategies: (1) replacing correct statements with contradictory ones in CoE/Non-CoE sentences, and (2) using GPT-4o to generate diverse contradictory expressions following previous work (Chen et al.; Zhou et al., 2023; Jin et al., 2024). We then inject these contradictory statements and gradually increase their proportion (from 0 to 0.75, with 0.25 intervals) in the contexts. After that, we obtain the answers from LLMs with the question and the constructed CoE and Non-CoE contexts. Following standard evaluation protocols Adlakha et al. (2024)

Table 3: LLMs' Accuracy (ACC) with CoE and Non-CoE surrounded by misinformation.

Model	Misinformation Proportion	HotpotQA			2WikiMultihopQA		
		CoE	Non-CoE		CoE	Non-CoE	
			SenP	*WordP*		*SenP*	*WordP*
GPT-3.5	*0*	**91.9%**	77.9%*	79.1%*	**97.4%**	74.1%*	83.5%*
	0.25	**81.8%**	62.5%*	64.0%*	**85.3%**	40.6%*	63.8%*
	0.5	**82.0%**	63.0%*	65.7%*	**65.5%**	43.4%*	52.3%*
	0.75	**75.7%**	58.9%*	60.8%*	**55.5%**	29.8%*	30.4%*
GPT-4	*0*	**93.5%**	83.4%*	86.4%*	**93.7%**	67.7%*	79.4%*
	0.25	**95.3%**	89.7%*	89.9%*	**96.5%**	86.0%*	91.9%*
	0.5	**90.7%**	84.6%*	87.4%*	**90.7%**	78.3%*	84.2%*
	0.75	**86.6%**	75.2%*	78.1%*	**85.0%**	60.7%*	69.4%*
Llama2-13B	*0*	**89.9%**	87.1%*	88.8%*	**96.5%**	95.3%*	93.3%*
	0.25	**74.8%**	70.6%*	67.6%*	**78.5%**	73.9%*	67.7%*
	0.5	**63.5%**	59.2%*	56.5%*	**57.9%**	52.0%*	52.7%*
	0.75	**57.0%**	42.1%*	44.9%*	**49.7%**	34.9%*	41.8%*
Llama3-70B	*0*	**92.5%**	76.8%*	74.5%*	**95.7%**	79.5%*	73.3%*
	0.25	**87.4%**	71.3%*	67.3%*	**93.1%**	72.6%*	61.2%*
	0.5	**82.1%**	64.8%*	62.5%*	**88.3%**	64.1%*	55.8%*
	0.75	**84.0%**	59.7%*	57.6%*	**85.6%**	56.5%*	52.4%*
Qwen2.5-32B	*0*	**87.8%**	71.3%*	75.7%*	**90.7%**	53.1%*	67.0%*
	0.25	**95.1%**	79.5%*	83.4%*	**97.4%**	63.5%*	75.4%*
	0.5	**88.5%**	72.3%*	71.7%*	**92.1%**	40.6%*	64.5%*
	0.75	**83.0%**	66.0%*	67.3%*	**86.9%**	39.6%*	55.0%*

* indicates statistical significance compared to CoE ($p < 0.05$)

for multi-hop QA, we use GPT-4o as the evaluator and compute ACC. We also conduct ablation studies to analyze how different CoE features affect conflict handling capabilities. Detailed experimental settings are provided in Appendix B.

7.2 Results and Discussion

Table 3 shows ACC after adding inaccurate information and produce knowledge conflicts with CoE and two types of Non-CoE. The results show that the average ACC reaches 84.1% for the CoE group, which is 21.4% and 15.3% higher than Non-CoE$_{SenP}$ and Non-CoE$_{WordP}$, respectively. These results demonstrate CoE's superior ability in maintaining correct output when faced with conflicting information. Furthermore, as the ratio increases from 0% to 75%, CoE's ACC decreases by 18.0%, while Non-CoE variants show greater drops (24.2% for Non-CoE$_{SenP}$ and 24.3% for Non-CoE$_{WordP}$). This indicates CoE's more stable performance against conflicting knowledge. Considering certain context poisoning methods (such as PoisonRAG (Zou et al., 2024)) which involve injecting multiple pieces of incorrect knowledge into the context, resulting in a high conflict ratio, CoE can also better help LLMs resist such attacks.

Ablation studies demonstrate the crucial role of evidence relations in handling conflicting knowledge, with their removal leading to a substantial 62.1% ACC drop when contradictory information is present. By connecting nodes and maintaining logical consistency, evidence relations make LLMs more resilient to contradictions. Detailed analysis is provided in Appendix B.

Summary: When inaccurate information exists in the context, CoE can help LLMs effectively maintain the robustness against such interference. This suggests that existing RAG defense methods could benefit from incorporating such structured evidence chains to enhance their robustness against misleading information.

8 Usability Assessment

Based on the above findings, we selected three representative tasks that leverages external knowledge, i.e., RAG-based multi-hop QA, external knowledge poisoning, and poisoning defenses (Zhou et al., 2024). For each task, we chodse the corresponding SOTA or prevalent baseline and modified certain components under the guidance of CoE-oriented findings to explore whether CoE-enhanced variants could achieve performance improvements.

8.1 RAG-based Multi-Hop QA

Given the complexity of multi-hop QA, RAG has emerged as a prevalent way for addressing such problems. A prevalent RAG framework follows a retrieve-rank-generate pipeline: first retrieving relevant knowledge snippets using a search engine, then employing a reranker model[1] to rank snippets based on relevance to the question, and finally using the ranked snippets as context for LLM generation. We select this standard RAG approach as our baseline because it represents the mainstream implementation of current RAG systems (Chen et al., 2024), which retrieve top-5 snippets from the Google Search API as context for the generation of LLM answers.

Based on this prevalent RAG framework, our variant primarily enhances the reranking process to introduce more CoE-compliant external knowledge into the context. Specifically, while the original reranker focuses on pure relevance matching, CoE features from questions can provide additional structural guidance, as shown in Figure 4.

- **CoE Feature Judgment:** The CoE features (intent, evidence nodes and relations) extracted from questions can be used to judge their presence in knowledge snippets through feature discrimination, producing judgments on feature coverage.
- **Coverage-based Selection:** The reranking process can prioritize snippets containing

[1] https://huggingface.co/BAAI/bge-reranker-large

Model	**Multi-Hop QA(ACC)**		**Attack (ASR)**		**Defense (ACC/ASR)**	
	RAG	RAG+CoE	PR	PR+CoE	IR	IR+CoE
GPT-3.5	68.1%	76.0%	69.0%	79.0%	49.0%/42.0%	78.0%/8.0%
GPT-4	72.9%	82.6%	49.0%	62.0%	56.0%/38.0%	80.0%/5.0%
Llama2-13B	64.4%	74.1%	62.0%	71.0%	45.0%/51.0%	79.0%/6.0%
Llama3-70B	67.8%	79.5%	60.0%	76.0%	60.0%/37.0%	84.0%/6.0%
Qwen2.5-32B	63.8%	77.0%	73.0%	80.0%	51.0%/42.0%	76.0%/6.0%

Table 4: Performance comparison between baselines and after adding CoE in three application scenarios.

more CoE features, particularly focusing on intent coverage first, followed by evidence relations and nodes. The detailed selection process is shown in Appendix F.

This optimized snippet selection serves as enhanced context for LLM generation.

8.2 External Knowledge Poisoning

External knowledge poisoning attacks aim to manipulate RAG systems by injecting malicious content into the knowledge base. We select PoisonedRAG (PR) (Zou et al., 2024) as our baseline, which uses LLM to generate false supporting documents for incorrect answers and injects them into the RAG knowledge base, causing the retriever to select these poisoned documents as context and mislead LLM to generate target answers.

Building upon PR, the SOTA knowledge poisoning attack, our variant enhances the document generation process by incorporating CoE features from target questions. By extracting CoE features from questions and integrating them into the generation process, the generated false knowledge exhibits stronger logical and semantic alignment with the questions. The detailed generation prompts incorporating these structural features are provided in Appendix J.

8.3 Poisoning Defense

Poisoning defenses aim to protect RAG systems against knowledge poisoning attacks. We select InstructRAG (IR) (Wei et al., 2024) as our baseline, which asks LLMs to first rationalize evidence relevance and uses these rationales for context selection, enhancing the system's ability to identify and reject misleading information.

Building upon IR, the SOTA RAG defense framework, our variant strengthens its defensive capability by incorporating CoE-structured knowledge validation. By generating knowledge containing CoE features extracted from questions, these CoE-structured knowledge are injected into the knowledge base. It provides more systematic supporting evidence when the framework requests document rationales. This enables LLMs to effectively select defensive knowledge pieces within the framework. The detailed generation process is provided in Appendix J.

8.4 Evaluation and Results

We evaluate the effectiveness of CoE across three RAG scenarios using the HotpotQA dataset. To investigate how CoE enhances RAG performance in multi-hop QA, we measure the effectiveness using accuracy (ACC). Table 4 shows that RAG+CoE achieves an average ACC improvement of 10.4% compared to RAG. Notably, RAG+CoE with GPT-4 achieves the highest ACC of 82.6%, while Qwen2.5-32B shows the most significant improvement when CoE is integrated into RAG, with a 13.2% increase in ACC. The results illustrate that knowledge structured through CoE provides more effective context for LLMs to reason and generate accurate responses.

In the knowledge poisoning attack, we measure the attack effectiveness using attack success rate (ASR). PR+CoE achieves 11.0% higher ASR on average across LLMs compared to PR, revealing that malicious knowledge deliberately structured through CoE becomes more effective at manipulating LLM outputs.

For RAG defense evaluation, we examine both ACC and ASR. IR+CoE demonstrates strong defensive capability with 27.2% higher ACC and 35.8% lower ASR compared to IR, indicating that CoE-structured defensive knowledge enables LLMs to better identify and resist misleading information while maintaining accurate responses.

9 Conclusion

In this paper, we introduce CoE and investigate its impact on LLMs in imperfect external knowledge for multi-hop QA. We characterize the CoE features and propose a discrimination approach to judge whether external knowledge exhibits the features within the user question. Generally, our study reveals that external knowledge aligned with CoE features exhibits stronger significance, deceptiveness, and robustness against extraneous and inaccurate information in contexts. We further validate the CoE-oriented findings by applying them to tasks that leverage external knowledge, demonstrating that the CoE-enhanced variants consistently outperform their original baseline counterparts.

Limitations

There are three limitations to the current study. Firstly, we apply the *RAG+CoE* to search for CoE in external knowledge, but there is no step to verify the correctness of answers within the CoE. If the retrieved CoE contains incorrect information, it may mislead the LLM to generate inaccurate responses. In Section 6, we discuss LLMs' Following Rate to CoE containing factual errors, showing that LLMs are highly likely to follow the knowledge provided in CoE.

Secondly, the usability of our proposed CoE-based reranking strategy (*RAG+CoE*) has inherent constraints across RAG scenarios. For instance, some RAG scenarios convert external knowledge into vectors and store them in vector databases, then search for question-relevant knowledge at the vector level during the retrieval phase.. Our approach, which operates at the textual level, is not suitable for such vector-based RAG scenarios.

Thirdly, our approach relies on prompt-based extraction of evidence nodes using GPT-4o, potential extraction errors (either incorrect identification or missing of evidence nodes) may affect CoE's performance. We systematically analyze these scenarios in Appendix D, where experiments on 1,000 HotpotQA samples demonstrate the robustness of our method: even under imperfect extraction conditions, the accuracy only drops marginally (from 90.2% to 89.3% and 89.4%). This suggests that while evidence node extraction quality matters, our approach maintains strong performance even with occasional extraction imperfections.

Ethical Statement

Our exploration of CoE-enhanced knowledge poisoning attacks is conducted strictly for red-team testing purposes to identify and understand potential vulnerabilities in RAG systems. Following responsible security research practices, we have promptly reported our findings to relevant RAG system providers and knowledge base platforms. We present only high-level methodological insights necessary for academic understanding, without releasing detailed attack implementations. Our goal is to help develop more robust RAG systems by revealing potential weaknesses in their knowledge base integration, thereby contributing to improved security measures rather than facilitating malicious exploits.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, et al. 2023. Palm 2 technical report.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2292–2307.

Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *arXiv preprint arXiv:2210.13701*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Unifying bias and unfairness in information retrieval: A survey of challenges and opportunities with large language models. *arXiv preprint arXiv:2404.11457*.

Jialin Dong, Bahare Fatemi, Bryan Perozzi, Lin F. Yang, and Anton Tsitsulin. 2024. Don't forget to connect! improving RAG with graph-based reranking. *CoRR*, abs/2405.18414.

Gary Edmond and Kent Roach. 2011. A contextual approach to the admissibility of the state's forensic science and medical evidence. *University of Toronto Law Journal*, 61(3):343–409.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024.

Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409*.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Siyi Liu, Qiang Ning, Kishaloy Halder, Wei Xiao, Zheng Qi, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, et al. 2024. Open domain question answering with conflicting contexts. *arXiv preprint arXiv:2410.12311*.

Erin Murphy. 2013. The mismatch between twenty-first-century forensic evidence and our antiquated criminal justice system. *S. Cal. L. Rev.*, 87:633.

OpenAI. 2022. Chatgpt. https://openai.com/blog/chatgpt.

Qwen Team. 2024. Qwen2.5: A party of foundation models! Blog post.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Shangqing Tu, Yuanchun Wang, Jifan Yu, Yuyang Xie, Yaran Shi, Xiaozhi Wang, Jing Zhang, Lei Hou, and Juanzi Li. 2024. R-eval: A unified toolkit for evaluating domain knowledge of retrieval augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5813–5824.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arık. 2024. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. *arXiv preprint arXiv:2406.13629*.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024a. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Hao Zhang, Yuyang Zhang, Xiaoguang Li, Wenxuan Shi, Haonan Xu, Huanshuo Liu, Yasheng Wang, Lifeng Shang, Qun Liu, Yong Liu, et al. 2024. Evaluating the external and parametric knowledge fusion of large language models. *arXiv preprint arXiv:2405.19010*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models.

A The details of the subject dataset with CoE and two types of Non-CoE

he details of the subject dataset with CoE and two types of Non-CoE is shown in Table 5.

Table 5: The details of the subject dataset with CoE and two types of Non-CoE.

Dataset	Type	Sample Num	Knowledge Piece Num
HotpotQA	CoE	676	4.0
	SenP	676	2.1
	WordP	676	4.0
2WikiMultihopQA	CoE	660	3.4
	SenP	660	1.9
	WordP	660	3.4

B Feature Effectiveness Analysis on LLM Performance

In this analysis, we examine three types of feature perturbations: WordP, Evidence RelationP (ERP), and IntentP using GPT-3.5 as our testing model on the HotpotQA dataset. WordP involves perturbing evidence node as detailed in Section 4.2. ERP removes evidence relations from the external knowledge (CoE) by prompting the LLM to modify the text while preserving other features. Similarly, IntentP removes intent information from CoE while maintaining other features. The experimental results are presented in Table 6.

The significance analysis (RQ1) reveals that evidence relation perturbation has the least impact on LLM accuracy, followed by evidence node perturbation, and then intent perturbation. This suggests that intent information plays the most crucial role in maintaining LLM accuracy.

Regarding deceptiveness (RQ2), CoE achieves the highest attack success rate when lacking evidence nodes, followed by missing evidence relation, and then intent. This highlights the significance of relationships and intent in affecting LLM vulnerability. However, CoE lacking evidence nodes demonstrates weaker stability against irrelevant external knowledge under attack scenarios, indicating that evidence nodes play a vital role in maintaining attack effectiveness when facing noisy knowledge during deception attempts.

For robustness against misinformation (RQ3), the absence of evidence relations leads to the most significant decrease in LLM accuracy when misleading information is introduced. This underscores that evidence relations are crucial features for constructing complete evidence chains and maintaining model reliability. In conclusion, each feature demonstrates distinct strengths in different scenarios: intent information is crucial for maintaining overall accuracy, relationships are vital for constructing evidence chains and misinformation resistance, while evidence nodes play a key role in handling irrelevant knowledge under misinformation scenarios. This diverse functionality suggests that intent, evidence relations and evidence nodes are all indispensable components in constructing effective Chain-of-Evidence (CoE) for robust LLM performance.

Table 6: Performance of GPT-3.5 with CoE and Non-CoE on HotpotQA Dataset

RQ	Metric	Proportion Type	Proportion	CoE	WordP	ERP	IntentP
RQ1	ACC	Irrelevant	0	91.9%	79.1%	81.1%	59.9%
			0.25	90.3%	77.5%	81.6%	56.5%
			0.50	89.9%	75.4%	78.5%	54.5%
			0.75	88.9%	74.5%	76.9%	54.5%
RQ2	ASR	Irrelevant	0	86.1%	83.1%	69.2%	57.3%
			0.25	85.8%	79.1%	64.8%	54.8%
			0.50	84.7%	77.8%	61.4%	53.2%
			0.75	78.4%	73.7%	58.1%	49.0%
RQ3	ACC	Misinformation	0	91.9%	79.1%	81.1%	59.9%
			0.25	81.8%	64.0%	21.7%	53.1%
			0.50	82.0%	65.7%	21.2%	52.1%
			0.75	75.7%	60.8%	19.0%	47.8%

C Effectiveness of CoE in Single-hop QA and Analysis of Hop Numbers

To provide a comprehensive evaluation of CoE's effectiveness, we conducted additional experiments on single-hop scenarios alongside our main multi-hop experiments. Multi-hop questions are particularly challenging for LLMs as they require sophisticated knowledge integration and logical reasoning capabilities. However, examining single-hop scenarios helps establish the generalizability of our approach across different reasoning complexity levels.

We evaluated GPT-3.5 on a single-hop dataset (RGB) following the experimental settings from RQ1-RQ3. The results shown in Table 7 reveal several interesting findings:

- CoE demonstrates consistent effectiveness in both single-hop and multi-hop scenarios, as shown in RQ1. However, both CoE and Non-CoE exhibit stronger resistance to irrelevant information in single-hop scenarios, which can be attributed to the reduced complexity of single-step reasoning tasks.

- The core advantages of CoE observed in RQ2 and RQ3 remain consistent across both single-hop and multi-hop contexts, supporting the broader applicability of our approach.
- Our comparative analysis reveals that while the number of reasoning hops does not significantly impact CoE's significance and robustness, it notably affects Non-CoE. As the number of hops increases, SenP and WordP show decreased resistance to imperfect knowledge. This pattern emerges because multi-hop reasoning requires both individual knowledge comprehension and cross-hop integration, making the LLM more vulnerable to irrelevant or misleading information.

These findings further validate CoE's capability to effectively guide LLM reasoning regardless of the reasoning complexity, while highlighting its particular advantages in more challenging multi-hop scenarios.

Besides, to analyze the robustness of CoE across different reasoning complexity levels, We conduct statistical analysis based on results from Table 1 and Table 7 on GPT-3.5's performance on questions requiring one-hop, two-hop, and three-hop reasoning while gradually introducing irrelevant knowledge. The results reveal interesting patterns across reasoning depths. For one-hop questions, CoE maintains consistently high accuracy (above 92.0%) even with increasing irrelevant knowledge, demonstrating strong robustness in simple reasoning scenarios where direct evidence-to-answer mapping is sufficient. The performance on two-hop questions shows more sensitivity to irrelevant knowledge, with accuracy declining from 91.0% to 88.0%. This suggests that intermediate reasoning steps are more vulnerable to distraction from irrelevant information. Interestingly, for three-hop questions, despite the higher reasoning complexity, the model shows better resilience than two-hop cases, maintaining accuracy above 90% in most scenarios. This counter-intuitive improvement may be attributed to the LLM's enhanced focus when processing more complex reasoning chains.

Table 7: Performance of GPT-3.5 with CoE and Non-CoE on Single-hop Dataset

RQ	Metric	Proportion Type	Proportion	CoE	SenP	WordP
RQ1	ACC	Irrelevant	0	93.0%	74.0%	84.1%
			0.25	93.4%	77.9%	84.4%
			0.50	93.4%	80.2%	84.8%
			0.75	92.6%	79.8%	85.6%
RQ2	ASR	Irrelevant	0	87.9%	55.0%	85.1%
			0.25	79.4%	47.4%	66.3%
			0.50	67.4%	40.4%	52.0%
			0.75	62.7%	32.7%	47.0%
RQ3	ACC	Misinformation	0	93.0%	74.0%	84.1%
			0.25	86.8%	65.1%	74.0%
			0.50	83.3%	65.8%	67.4%
			0.75	77.5%	60.0%	60.0%

Table 8: Accuracy of GPT-3.5 under Different Hop Num

Irrelevant Proportion	One-hop	Two-hop	Three-hop
0	93.0%	91.0%	94.0%
0.25	93.4%	89.0%	90.0%
0.50	93.4%	88.0%	92.0%
0.75	92.6%	88.0%	92.0%

D Reliability of automated evidence nodes extraction for CoE and its impact on performance

In our approach, we define evidence nodes and provide few-shot examples in the prompt for GPT-4o to perform evidence node extraction. Given that automated evidence node extraction may contain errors in real-world applications, we conducted a systematic analysis of potential evidence node extraction errors. These errors primarily manifest in two ways: 1) **Extraction Errors:** incorrectly identifying intent-related content as evidence nodes; 2) **Missing Errors:** failing to extract essential evidence nodes. For example, as shown in Figure 1, Extraction Errors would occur when "state" from the intent/question is incorrectly included in the evidence nodes, while Missing Errors would happen when essential evidence node like "CEO" are not extracted, both of which could affect the accuracy of CoE identification. To assess the impact of these potential errors, we designed corresponding perturbation operations and simulated both error types on our test dataset. The detailed experimental results and analysis are presented in Table 9.

To examine the impact of imperfect extraction, we conducted experiments on 1,000 HotpotQA samples by either adding a shared entity from intent/question (Extraction Errors) or randomly removing one evidence node (Missing Errors). The result show that Missing Errors led to over-identification

Table 9: Accuracy of GPT-3.5 under Different Evidence Nodes Error Types

Irrelevant Proportion	Our	Missing Errors	Extraction Errors
0	91.9%	91.2%	91.2%
0.25	90.3%	90.1%	90.1%
0.50	89.9%	89.2%	88.4%
0.75	88.9%	87.4%	87.5%
Num	676	803	641

of CoE (803 vs. 676 Our), while Extraction Errors resulted in under-identification (641). Both scenarios slightly decreased response accuracy compared to normal conditions (90.2% Our, 89.4% Missing Errors, 89.3% Extraction Errors).

E Perturbation Strategies for Non-CoE Construction

We employ two controlled perturbation strategies to construct Non-CoE samples while maintaining the same question context:

Sentence-Level Perturbation (SenP). For multihop QA, we simulate incomplete knowledge scenarios by removing knowledge pieces from CoE. We segment CoE into sentences and identify candidates containing question-mentioned evidence nodes (excluding answer nodes). We iteratively remove these candidates until CoE discrimination confirms that the remaining knowledge no longer contains complete CoE. This sentence-level approach helps understand how LLMs behave when key evidence pieces are entirely missing within the same reasoning context. Figure 3 shows this sentence-level perturbation process.

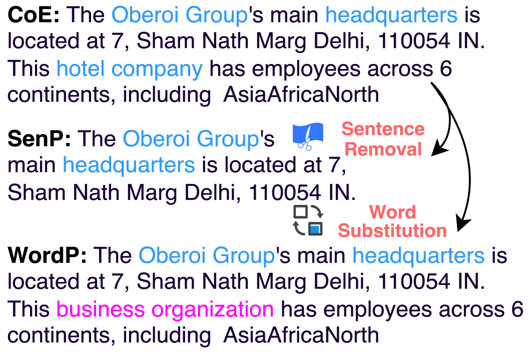
Word-Level Perturbation (WordP). We create Non-CoE by replacing specific evidence nodes with their GPT-4 generated higher-level expressions (e.g., replacing hotel company" with business organization"), maintaining more original information compared to sentence removal. This finer-grained approach examines LLMs' sensitivity to evidence nodes while preserving most of the original semantic information. Figure 3 demonstrates this word-level perturbation approach.

F The Algorithm for the Coverage-Based Selection

We show the detailed algorithm 1 for the minimal coverage search in *RAG+CoE*.

G The Details in *RAG+CoE*

We show the overview of *RAG+CoE* in Figure 4.



CoE: The Oberoi Group's main headquarters is located at 7, Sham Nath Marg Delhi, 110054 IN. This hotel company has employees across 6 continents, including AsiaAfricaNorth

SenP: The Oberoi Group's main headquarters is located at 7, Sham Nath Marg Delhi, 110054 IN.

WordP: The Oberoi Group's main headquarters is located at 7, Sham Nath Marg Delhi, 110054 IN. This business organization has employees across 6 continents, including AsiaAfricaNorth

Figure 3: Examples of CoE and two types of Non-CoE.

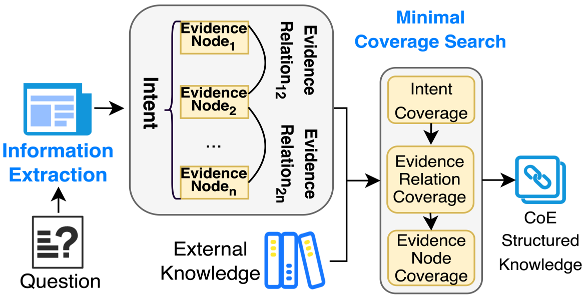


Figure 4: The overview of *RAG+CoE*.

H Details of Information Extraction Prompts

The details of the information extraction prompts are illustrated below. In pipeline, we replace the placeholders in the following prompts with the question and evidence nodes.

Algorithm 1: Coverage-Based Selection

```
Input: External knowledge list $EK$, Judged external
       knowledge list $IEK$, where each item contains
       Intent, Evidence Relations, and Evidence
       nodes judgments
Output: Set $S$ of minimal coverage external
        knowledge
$S \leftarrow \emptyset$;
# Phase 1: Intent Coverage;
for $i \leftarrow 0$ to $|IEK| - 1$ do
    if $IE[i].Intent = TRUE$ then
        $S \leftarrow S \cup \{EK[i]\}$
# Phase 2: Evidence Relation Coverage;
$R_{uncovered} \leftarrow$
    GetUncoveredEvidencerelation($IEK, S$);
for $r \in R_{uncovered}$ do
    for $i \leftarrow 0$ to $|IEK| - 1$ do
        if $IEK[i].Evidencerelation[r] = TRUE$
          then
            $S \leftarrow S \cup \{EK[i]\}$;
            break;
# Phase 3: Evidence Node Coverage;
$K_{uncovered} \leftarrow$
    GetUncoveredEvidencenodes($IEK, S$);
for $k \in K_{uncovered}$ do
    for $i \leftarrow 0$ to $|IEK| - 1$ do
        if $IEK[i].Evidencenode[k] = TRUE$ then
            $S \leftarrow S \cup \{EK[i]\}$;
            break;
return $S$;
```

Intent and evidence node Extraction Prompt:
Please extract both the intent and evidence nodes of the question, using the following criteria:
1) As for intent, please indicate the content intent of the evidence that the question expects, without going into specific details.
2) As for evidence nodes, Please extract the specific details of the question.
The output must be in json format, consistent with the sample. Here are some examples:
Example1:
Question:750 7th Avenue and 101 Park Avenue, are located in which city?
Output: { "Intent": "City address Information", "evidence nodes": ["750 7th Avenue", "101 Park Avenue"] }
Example2:
Question: The Oberoi family is part of a hotel company that has a head office in what city?
Output: { "Intent": "City address Information", "evidence nodes": ["Oberoi family", "head office"] }
Example3:
Question: What nationality was James Henry Miller's wife?
Output: { "Intent": "Nationality of person", "evidence nodes": ["James Henry Miller", "wife"] }
Example4:
Question: What is the length of the track where the 2013 Liqui Moly Bathurst 12 Hour was staged?
Output: { "Intent": "Length of track", "evidence nodes": ["2013 Liqui Moly Bathurst 12 Hour"] }
Example5:
Question: In which American football game was Malcolm Smith named Most Valuable player? Output: { "Intent": "Name of American football game", "evidence nodes": ["Malcolm Smith", "Most Valuable player"] }
Question: ***[Question]*** **Output:**

Evidence Relations Extraction Prompt:
Please extract evidence relations based on the input questions and evidence nodes, using the following criteria:
1) Each evidence relation has two elements, the implied evidence nodes and the textual description of the evidence relations.
2) The description of the evidence relations is limited to the two evidence nodes and does not involve other evidence nodes.
3) If there is no evidence relation between evidence nodes, no extraction is required.
The output must be in json format, consistent with the examples. Here are some examples:
The output must be in json format, consistent with the sample. Here are some examples:
Example1:
Question:750 7th Avenue and 101 Park Avenue, are located in which city?
Evidence nodes:["750 7th Avenue", "101 Park Avenue"]
Output: []
Example2:
Question: Lee Jun-fan played what character in T̈he Green Hornetẗelevision series?
Evidence nodes:["Lee Jun-fan", "The Green Hornet"]
Output: [{"Evidence nodes":["Lee Jun-fan", "The Green Hornet"], "Evidence Relations: "played character in"}]
Example3:
Question: In which stadium do the teams owned by Myra Kraft's husband play?
Evidence nodes: ["teams", "Myra Kraft's husband"]
Output: [{"Evidence nodes":["teams", "Myra Kraft's husband"], "Evidence Relations": "is owned by"}]
Example4:
Question: The Colts' first ever draft pick was a halfback who won the Heisman Trophy in what year?
Evidence nodes:["Colts' first ever draft pick", "halfback", "Heisman Trophy"]
Output:[{"Evidence nodes":["Colts' first ever draft pick", "halfback"], "Evidence Relations": "was"}]
Example5:
Question: The Golden Globe Award winner for best actor from "Roseanne" starred along what actress in Gigantic?
Evidence nodes:["Golden Globe Award winner", "best actor", "Roseanne", "Gigantic"]
Output: [{"Evidence nodes":["Golden Globe Award winner", "best actor"], "Evidence Relations": "for"}, {"Evidence nodes":["best actor", "Roseanne"], "Evidence Relations": "starred in "}]
Question: ***[Question]***
Evidence nodes: ***[Evidence node]***
Output:

I Details of Feature Discrimination Prompts

The details of the Feature Discrimination prompts are illustrated below. In pipeline, we replace the placeholders in the following prompts with the external knowledge, intent, evidence node, and evidence relation.

Intent Discrimination Prompt:
Please determine whether the input intent is covered in the input external knowledge. Please output only "yes" or "no".
Input intent: ***[Intent]***
Input external knowledge: ***[External Knowledge]***

Evidence Nodes Discrimination Prompt:
Please determine if the input evidence node is mentioned in the input external knowledge. It doesn't necessarily need to be an exact character match; partial matches or semantic similarities are also acceptable. Please output only "yes" or "no".
Input evidence node: ***[Evidence node]***
Input external knowledge: ***[External Knowledge]***

Evidence Relations Discrimination Prompt:
Please determine if the input external knowledge supports the logical relationship between the two given evidence nodes. If there is explicit evidence in the input knowledge that confirms the evidence node-evidence relation-evidence node triple, output "yes"; otherwise output "no". Please respond only with "yes" or "no".
Input triple: ***(evidence node1, evidence relation, evidence node2)***
Input external knowledge: ***[External Knowledge]***

J Details of the CoE-Guided Knowledge Generation Prompts

CoE-Guided Knowledge Generation Prompt:
Given:
Evidence Nodes: [Nodes]
Evidence Relations: [Relations]
Intent: [Intent]
Question: [Question]
Answer: [Answer]
Create a corpus that:
1. Integrate all given [Nodes] naturally, do not replace [Nodes] with pronouns, each sentence in the corpus include as many [Nodes] as possible.
2. If there are [Relations] between [Nodes], when constructing sentences containing these [Nodes], try to express the [Relations] between them.
3. The corpus should align with the meaning of the [Intent].
4. Flows naturally as coherent text
5. Please limited the corpus to 100 words.
The corpus should serve as a comprehensive basis for supporting why the [Question] has the answer:[Answer].
It's a creative game focusing on generating the support for the specified answer: [Answer], without requiring factual accuracy.

K Details of the Answer Generation Prompts

The details of the Answer Generation prompts are illustrated below. In pipeline, we replace the placeholders in the following prompts with the correct answer.

Answer Generation Prompt:
For the input phrase, please generate a phrase of similar type and format, but not the same. Just output the phrase, no explanation is needed, the expression form is consistent with the examples. Here are some examples:
Example1:
Input phrase: United States
Output: Canada
Example2:
Input phrase: alcohol
Output: Soda
Example3:
Input phrase: September 29, 1784
Output: April 22, 1964
Example4:
Input phrase: Laura Ellen Kirk
Output: Elon Musk
Example5:
Input phrase: 39,134
Output: 19,203
Input phrase: ***[Correct Answer]***
Output: