What External Knowledge is Preferred by LLMs? Characterizing and Exploring Chain of Evidence in Imperfect Context

Anonymous ACL submission

Abstract

Incorporating external knowledge into large language models (LLMs) has emerged as a promising approach to mitigate outdated knowledge and hallucination in LLMs. However, external knowledge is often imperfect. In addition to useful knowledge, external knowledge is rich in irrelevant or misinformation in the context that can impair the reliability of LLM responses. This paper focuses on LLMs' preferred external knowledge in imperfect contexts when handling multi-hop QA. Inspired by criminal procedural law's Chain of Evidence (CoE), we characterize that knowledge preferred by LLMs should maintain both relevance to the question and mutual support among knowledge pieces. Accordingly, we propose an automated CoE discrimination approach and evaluate LLMs' effectiveness, faithfulness and robustness with CoE, including its application in the knowledge-intensive task. Tests on five LLMs show CoE improves generation accuracy, answer faithfulness, robustness to knowledge conflicts, and performance in a knowledge-intensive task.

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knowledge acquired by large language models (LLMs) through pre-training at a specific point in time becomes outdated with the knowledge evolution or produces hallucination (Achiam et al., 2023; Touvron et al., 2023a; Anil et al., 2023).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 useful knowledge that users expect LLMs to follow (as shown in Figure 1), the context typically contains two types of noise (Chen et al., 2024; Zou et al., 2024): 1) Irrelevant information, despite showing textual similarities with the question, cannot support the correct answer (Chen et al., 2024; Xiang et al., 2024); 2) Misinformation, which can confuse LLMs and lead to incorrect answers (Liu et al., 2024). Especially when dealing with complex scenarios like multi-hop QA, the acquisition of such noise is inevitable due to limitations of retrievers or quality deficiencies in the specialized knowledge corpus (Wang et al., 2024; Shao et al., 2024; Dai et al., 2024; Tang and Yang, 2024). This hinders LLMs from effectively utilizing useful knowledge within external knowledge and leads to incorrect answers.

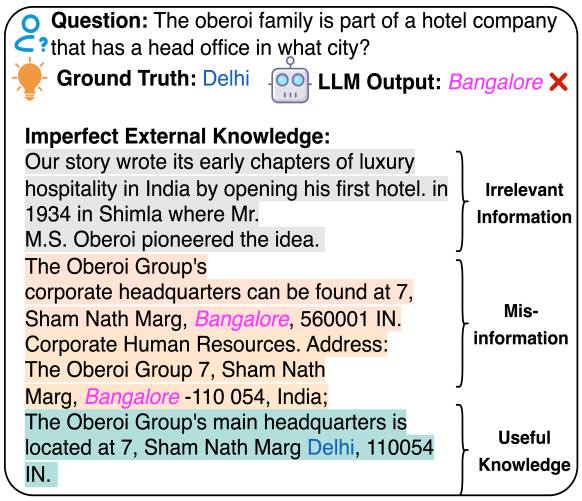

Figure 1: Example of imperfect external knowledge.

To this end, many studies focus on investigating the external knowledge preferences of LLMs in imperfect context (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 have mainly the following two deficiencies: 1) They focus on qualitative findings and lack automated discrimination given external knowledge, such as it is promising to determine whether external knowledge meets the completeness criteria in completeness bias (Zhang

et al., 2024); 2) They focus on single-hop QA, where a single piece of knowledge can cover all the necessary elements for QA, and whether the findings hold in complex scenarios is unclear.

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. Analogously to the scenario where LLMs rely on external knowledge for QA, we consider that the preferred knowledge should show relevance to the question (relevance) and mutual support and complementarity among knowledge 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 contains CoE. After that, we investigate the LLMs' preference towards CoE from four aspects below.

- **Effectiveness** where we investigate whether LLMs perform better when external knowledge contains CoE compared to the situation where it contains relevant information but does not constitute a CoE.
- **Faithfulness** where we extremely set the CoE's answer to be incorrect and observe LLMs' adherence even when the CoE contains factual errors.
- **Robustness** where we explore whether CoE can help improve the resistance of LLM to external knowledge occupied by misinformation which results in the knowledge conflicting.
- **Usability** where we select a knowledge-intensive task, specifically multi-hop question answering, and design a CoE-guided retrieval strategy to explore the effectiveness.

Using HotpotQA (Yang et al., 2018) and 2WikiMultihopQA (Ho et al., 2020) as sources, we constructed 1,336 multi-hop QA pairs and the corresponding CoE based on the proposed CoE discrimination approach. 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) and obtain the following main findings.

- External knowledge equipped with CoE can more effectively (than Non-CoE) help LLMs generate correct answers in context rich with irrelevant information.
- LLMs exhibit higher faithfulness to the answer implicated in CoE (than Non-CoE), even when CoE contains factual errors.
- LLMs exhibit higher robustness against knowledge conflict (than Non-CoE) if the external knowledge is equipped with CoE.
- The CoE-guided retrieval strategy can effectively improve LLM's accuracy in knowledge-intensive task.

The above findings could provide insights for future research in designing the retrieval process and assessing the quality of external knowledge with the proposed CoE discrimination approach. Furthermore, the content safety of CoE should also be a concern considering the faithfulness, as adversaries can also exploit CoE to generate targeted manipulations. 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 Discrimination Approach

3.1 CoE Characterization

Drawing from the law of criminal procedure, judicial decisions in cases require the formation of a Chain of Evidence (CoE) through evidence collection (Edmond and Roach, 2011; Murphy, 2013). Such evidence must demonstrate two properties: relevance (pertaining to the case) and interconnectivity (evidence mutually supporting each other). We analogize judicial decisions to the scenario in which LLMs identify correct answers from external knowledge in response to input questions.

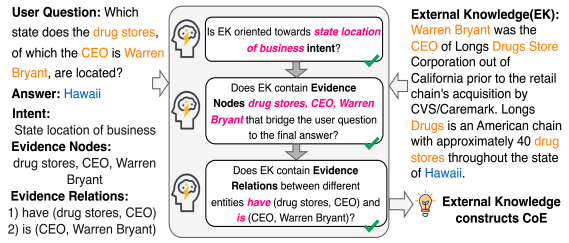

Figure 2: Example of CoE and the CoE features.

We assume that LLMs prefer external knowledge forming CoE, and characterize three features based on CoE's required two properties.

- **Intent** is a noun or noun phrase that describes the final answer a user intends to solve through the 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 within the external knowledge, corresponding to the essential knowledge elements to reason from a user question to the final answer. It ensures the consistency between external knowledge and the user question from the entity perspective.
- **Evidence Relations** are logical predicates within external knowledge, 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 user question.

Taking Figure 2 as an example, intent specifies "state location of business" as the user question goal, indicating the user wants to find the state where the business operates. Evidence nodes are the key entities extracted from user question, i.e., "drug stores", "CEO", and "Warren Bryant". These nodes serve as bridges to connect the question with external knowledge about "Longs Drugs Store Corporation". 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 evidence chain that forms a complete knowledge structure tailored to the specific question.

3.2 CoE Discrimination Approach

Based on the characterized features, we design an approach to discriminate whether external knowledge qualifies as CoE, as illustrated in Figure 5. First, for each question, we perform information extraction to extract its inherent intent, evidence nodes, and evidence relations. Based on GPT-4o, we adopt the prompt used in the previous study (Li et al., 2023) and enhance it by few-shot learning (adding 5 extra input-output samples) to help LLM achieve better extraction performance. Appendix G shows the template for the extraction prompt.

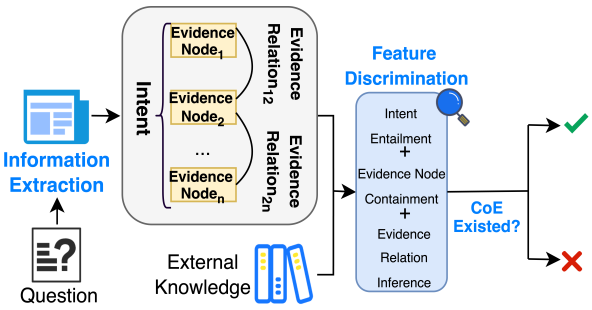

Figure 3: The overview of CoE discrimination approach.

Second, for external knowledge, the pipeline discriminates whether it contains CoE. Specifically,

the approach leverages GPT-4o to discriminate the presence of intent, evidence nodes, and evidence relations within external knowledge. As for intent, analogous to the textual entailment task, LLMs treat external knowledge as a premise and intent as a hypothesis, reasoning whether the hypothesis holds based on the given premise. For evidence nodes, the LLM identifies phrases contained in external knowledge that are semantically similar with evidence nodes. For evidence relations, the LLM employs its logical reasoning capabilities to identify and establish connections between evidence nodes. External knowledge is discriminated as CoE exists if all extracted features is present, and as CoE does not exist if any feature is missing. The prompts for feature discrimination are provided in the Appendix H.

4 Subject Dataset and LLMs

4.1 CoE Sample Construction

We selected two commonly used multihop QA datasets, HotpotQA and 2WikiMultihopQA as the sample sources. In the two datasets, each sample consists of a question, an answer, and supporting knowledge to derive the answer to each question. Given the nature of multi-hop QA, supporting knowledge typically contains multiple knowledge pieces[1]. Since this knowledge bridges questions to answers, it likely exhibits CoE features. Therefore, we consider it as a candidate CoE for each QA pair.

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 candidates contain CoEs. Finally, we obtained 676 and 660 samples that contain CoE from candidates, with an average of 4.0 and 3.4 knowledge pieces for two datasets, respectively (details in Table 1).

Table 1: 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

[1]A knowledge piece refers to a complete sentence

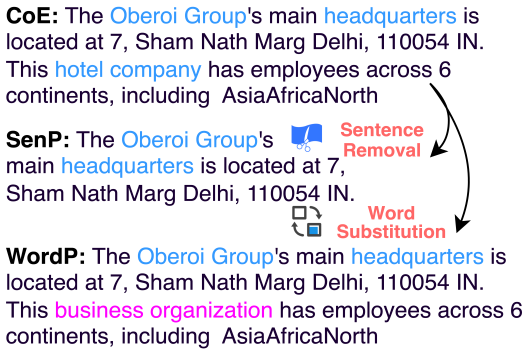

Figure 4: Examples of CoE and two types of Non-CoE.

4.2 Non-CoE Sample Construction

Based on the CoE samples, we construct Non-CoE samples where knowledge pieces fail to satisfy either the relevance or interconnectivity property of CoE. During the process, two strategies are utilized.

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 the remaining knowledge no longer contains complete CoE. Figure 4 shows this sentence-level perturbation process.

Word-Level Perturbation (WordP). We create Non-CoE by replacing specific evidence nodes with their GPT-4o generated higher-level expressions (e.g., replacing "hotel company" with "business organization"), maintaining more original information compared to sentence removal. Figure 4 demonstrates this word-level perturbation approach.

Finally, for each QA pair, we construct a five-element tuple, *<Question, Answer, CoE, SenP, WordP>*, while SenP and WordP serve as comparison samples, form the basis for subsequent experiments (details in Table 1).

4.3 Studied LLMs

For the following experimantal evaluation, we introduce two closed-source LLMs (GPT-3.5, GPT-4) and three open-source LLMs (LLama2-13B, LLama3-70B, and Qwen2.5-32B). All subsequent experiments are evaluated across these LLMs.

5 Effectiveness Assessment

Starting from the constructed CoE and Non-CoE samples, we inject additional irrelevant pieces into their contexts and investigate whether CoE can

Table 2: 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 ($p < 0.05$)

better help LLMs generate correct answers under external information rich with irrelevant noise.

5.1 Experimental Setup

For each sample <Question, Answer, CoE, SenP, WordP>, we collect irrelevant information by searching Google with evidence nodes from the Question using the template "Please introduce the background of [evidence node]". This ensures retrieved content is lexically similar but semantically irrelevant. We then inject this irrelevant information into "CoE", "SenP", and "WordP" at four different ratios (0.25 intervals) based on character length. Finally, we input the "Question" with "CoE", "SenP", and "WordP" to the LLMs for evaluation.

For each sample, we evaluate the consistency between LLM outputs and ground truth "Answer" using GPT-4o as the judge, following the evaluation method in Adlakha et al. (2024). We calculate the accuracy (ACC) for each LLM across three groups ("CoE", "SenP", "WordP").

5.2 Results and Findings

Table 2 shows the response accuracy of LLMs using CoE and two types of Non-CoE under different proportions of irrelevant information. The main findings and supporting results are illustrated below.

Finding-1: External knowledge equipped with CoE can help LLMs generate correct answers more effectively than Non-CoE. Generally, experimental results show that CoE achieves an average accuracy of 92.0% across five LLMs and two datasets, outperforming Non-CoE variants SenP and WordP by 22.5% and 16.3%, respectively. Moreover, compared to CoE, we conducted Mann-Whitney tests (Mann and Whitney, 1947) on all experiment groups. The results of the hypothesis test show that the improvement in CoE across all types of Non-CoE is statistically significant (significant level is 0.05).

Finding-2: LLMs exhibit greater resistance if CoE exists in external knowledge as the proportion of irrelevant information increases. As the proportion of irrelevant increases from 0% to 75%, the ACC of LLMs with CoE only decreases by 1.8%, while the ACC decreases by 12.9% and 9.0% under the Non-CoE variants SenP and WordP, respectively. WordP outperforms SenP in Non-CoE scenarios with higher accuracy and better resilience to irrelevant information. The superior performance of WordP's richer content suggests that information density benefits LLM QA capabilities. Yet CoE still achieves better results than WordP despite similar information content, emphasizing the importance of complete evidence chains.

We analyzed the impact of different CoE features and reasoning hops on LLM effectiveness. The experiments show that intent information has the strongest influence on LLM accuracy, followed by evidence relations and nodes; meanwhile, under increasing irrelevant knowledge, single-hop reasoning maintains the most stable performance (>92.0% ACC), followed by three-hop (>90.0% ACC), while two-hop reasoning shows higher sensitivity with ACC dropping from 91.0% to 88.0%. Detailed experimental results and analysis are provided in Appendix A and B.

6 Faithfulness Assessment

Based on the effectiveness assessment, we investigate a more challenging scenario, where the CoE contains factual errors, to determine whether LLMs can still exhibit a certain degree of faithfulness and produce answers consistent with the incorrect answer in CoE.

6.1 Experimental Setup

For the five-element tuple (<*Question, Answer, CoE, SenP, WordP*>), we respectively substitute the correct answers in "CoE", "SenP" and "WordP" with the incorrect ones to simulate the relevant knowledge contains the factual errors. To maintain textual coherence after the answer substitution, we construct incorrect answers that match the original

Table 3: LLMs' Following Rate (FR) on CoE and Non-CoE.

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 ($p < 0.05$)

in both type and format. We generate incorrect answers by replacing evidence nodes with those of the same type (e.g., "United States" to "Canada") and format (e.g., dates), using GPT-4 for answer type classification. Manual inspection confirms 100.0% type and format consistency between generated and correct answers. The detailed prompt design is provided in Appendix I.

To investigate LLMs' faithfulness with CoE under imperfect external knowledge, we progressively add irrelevant information to the external knowledge. The specific process follows the same procedure as described in Section 5.1. As for the evaluation metric, we use Following Rate (FR), defined as the proportion of all the LLM outputs consistent with incorrect answers contained in "CoE", "SenP" or "WordP" respectively. Following the previous study Adlakha et al. (2024), GPT-4o is used to evaluate consistency.

6.2 Results and Findings

Table 3 shows the FR of LLMs with external knowledge under CoE and two types of Non-CoE containing incorrect answers. The main findings and supporting results are illustrated in the following.

Finding-3: LLMs exhibit significant faithfulness to the answer supported by CoE although it contains factual errors. The results show that under CoE, the average FR reaches 85.4%, which is 20.6% and 16.2% higher than the SenP and WordP types under Non-CoE respectively. Moreover, Mann-Whitney tests confirmed statistically significant improvements of CoE over all Non-CoE groups ($p < 0.05$).

Finding-4: LLMs following CoE demonstrate higher stability against irrelevant noise variations when handling factual errors, compared to Non-CoE. As irrelevant information in external knowledge increases from 0% to 75%, the FR of LLMs with CoE decreases by 3.6%, while the FR drops by 9.7% and 7.9% under Non-CoE variants SenP and WordP, respectively.

Beyond the main findings, LLMs show 6.6% lower FR when handling incorrect CoE versus correct CoE (Table 2), suggesting their parametric knowledge helps detect and correct some factual errors. We analyze how different CoE features affect LLM faithfulness, finding that while knowledge lacking evidence nodes shows highest compliance, it exhibits weaker resistance to irrelevant knowledge under misinformation scenarios. Detailed analysis is provided in Appendix A.

7 Robustness Assessment

We make the knowledge conflicts by injecting the misinformation in the context of CoE and Non-CoE. Robustness explores whether CoE can help LLMs more effectively resist the conflict and produce the correct answers.

7.1 Experimental Setup

We generate misinformation that contains factual errors and conflicts with CoE/Non-CoE knowledge using two strategies: (1) replacing correct answers with incorrect ones in CoE sentences, and (2) using GPT-4o to generate multiple incorrect answer expressions, following previous work (Chen et al.; Zhou et al., 2023; Jin et al., 2024).

To investigate how CoE affects LLM performance as the proportion of misinformation increases, we continuously increase the proportion of misinformation and inject it into the context of CoE and Non-CoE respectively. After injection, since there are both correct and incorrect statements of the same subject within the external knowledge, leading to the knowledge conflict. Then, we send questions and conflicting external knowledge to the LLMs and assess their performance using ACC.

7.2 Results and Findings

Table 4 shows LLMs' response accuracy (ACC) after adding misinformation to CoE and two types

Table 4: 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 ($p < 0.05$)

of Non-CoE. The main findings and supporting results are illustrated in the following.

Finding-5: LLMs augmented with CoE exhibit higher robustness against knowledge conflict than Non-CoE. The results show that under CoE, the average ACC of LLMs reaches 84.1%, which is 21.4% and 15.3% higher than the SenP and WordP types under Non-CoE respectively. Besides, as the proportion of misinformation increases from 0% to 75%, LLMs' ACC under CoE shows 6.2% and 6.3% smaller decreases compared to the reductions observed in SenP and WordP under Non-CoE.

Finding-6: Compared to adding irrelevant information to CoE, adding misinformation has a greater impact on LLM's ability to generate correct outputs. In Table 2, when adding irrelevant information from 0% to 75%, the ACC of LLMs with CoE only decreases by 1.8%. However, as shown in Table 4, introducing misinformation under similar settings results in an 18.0% ACC drop for LLMs equipped with CoE. We analyze the contribution of different CoE features to LLM robustness, revealing that evidence relations are most crucial for misinformation resistance, with their absence causing the largest accuracy drop when misleading information is introduced. Detailed analysis is provided in Appendix A.

8 Usability Assessment

We selected a knowledge-intensive task for the case analysis and designed a CoE-guided retrieval strategy to investigate the extent to which CoE improves performance compared with baseline approaches.

8.1 CoE-guided Retrieval Strategy

We design a retrieval strategy (*ScopeCoE*) guided by CoE. For a given question, *ScopeCoE* first employs a search engine to retrieve an initial pool of relevant knowledge snippets. Then *ScopeCoE* selects the minimal set of knowledge snippets that encompass a CoE. It consists of two phases: 1) **CoE Feature Judgment**, which judges the CoE features covered by each knowledge snippet; 2) **Minimal Coverage Search**, which finds the minimal set of knowledge snippets that cover CoE. The overview of *ScopeCoE* is shown in Appendix F.

8.1.1 CoE Feature Judgment

ScopeCoE first extracts CoE features from the question and then judges them in each knowledge snippet. Specifically, as shown in Figure 5, *ScopeCoE* employs the same information extraction component in the discrimination approach to extract the intent, evidence nodes and evidence relations from the question. Then, for each knowledge snippet, *ScopeCoE* utilizes the proposed feature discrimination approach to determine whether it contains these extracted features, and records the judgment results. Finally, we obtain a set of judgments regarding intent, evidence nodes, and evidence relations for each knowledge snippet.

8.1.2 Minimal Coverage Search

After obtaining the judgment set, *ScopeCoE* searches for the minimal set of textual snippets that cover CoE. The algorithm process is shown in Appendix E. First, *ScopeCoE* searches for knowledge snippets that contain intent and adds them to the minimal set. Second, *ScopeCoE* examines the coverage of the evidence relations. Specifically, it determines whether the minimal set already contains all evidence relations. If there are uncovered evidence relations, it searches the remaining knowledge snippets and adds those containing uncovered evidence relations to the minimal set. Finally, *ScopeCoE* proceeds to examine evidence nodes coverage following the same process. It checks if the minimal set covers all evidence nodes. If uncovered evidence nodes exist, it searches the remaining snippets for those containing these evidence nodes.

ScopeCoE manages to search for the minimal set that completely covers all CoE features, ultimately outputting a set of knowledge snippets that covers the maximum number of CoE features, which serves as context input for the LLM.

Table 5: LLMs' Accuracy (ACC) on baselines and *ScopeCoE*.

Scenarios	Model	HotpotQA				2WikiMultihopQA			
		CoT_SC	*VE*	*RAG*	*ScopeCoE*	*CoT_SC*	*VE*	*RAG*	*ScopeCoE*
KI	GPT-3.5	19.7%	**33.7%**	30.2%	31.6%	9.3%	19.0%	19.9%	**23.9%**
	GPT-4	37.9%	41.5%	40.5%	**43.0%**	30.8%	32.3%	32.1%	**36.0%**
	Llama2-13B	8.3%	29.7%	28.6%	**32.9%**	1.8%	16.7%	15.8%	**21.1%**
	Llama3-70B	27.0%	32.9%	32.4%	**35.6%**	4.6%	19.9%	19.2%	**22.7%**
	Qwen2.5-32B	17.9%	**35.7%**	30.6%	31.6%	4.2%	**23.1%**	20.7%	21.1%
KC	GPT-3.5	19.7%	33.7%	68.1%	**76.0%**	9.3%	19.0%	54.6%	**81.5%**
	GPT-4	37.9%	41.5%	72.9%	**82.6%**	30.8%	32.3%	59.3%	**88.6%**
	Llama2-13B	8.3%	29.7%	64.4%	**74.1%**	1.8%	16.7%	51.7%	**74.0%**
	Llama3-70B	27.0%	32.9%	67.8%	**79.5%**	4.6%	19.9%	49.4%	**80.0%**
	Qwen2.5-32B	17.9%	35.7%	63.8%	**77.0%**	4.2%	23.1%	49.4%	**83.8%**

8.2 Experimental Setup

We used the constructed CoE samples (including "Question", "Answer" and "CoE") for usability evaluation. We evaluate our method under two scenarios: 1) **Knowledge Incomplete (KI)** represents real-world scenarios where knowledge retrieval solely depends on Google Search API, simulating practical situations where we can only access publicly available information through search engines. Such retrieved knowledge is often insufficient to fully answer questions. 2) **Knowledge Complete (KC)** refers to scenarios where the search corpus contains sufficient knowledge to form complete evidence chains. In this setting, we first use Google Search API to retrieve relevant snippets for each question, then augment the corpus by decomposing CoE into multiple knowledge pieces based on sentence completeness.

We compare *ScopeCoE* with three baselines: 1) *RAG (Chen et al., 2024)* retrieves top-5 most relevant snippets from the external corpus as context for LLMs' generation; 2) *CoT-SC (Wang et al., 2023)* enhances Chain-of-Thought reasoning by sampling diverse rationales and selecting the most consistent answers rather than using naive greedy decoding; 3) *VE (Zhao et al., 2023)* is a state-of-the-art framework that improves prediction factuality by post-editing CoT rationales with external knowledge. We evaluate methods on both effectiveness (Accuracy) and efficiency (Number of LLM calls). For efficiency measurement, we count the additional LLM calls required beyond the base retrieval process, as this directly reflects the computational overhead of different approaches.

8.3 Results and Findings

Finding-7: CoE guided reasoning can effectively leverage partial information for intermediate reasoning steps, making it robust in scenarios with incomplete knowledge retrieval. In the knowledge incomplete (KI) scenario, where only the Google Search API is employed for information retrieval, *ScopeCoE* demonstrates notable performance gains on both HotpotQA and 2WikiMultihopQA. Specifically, it achieves average ACC of 34.9% and 25.0%, respectively, surpassing established baselines such as CoT_SC and VE by 0.2%–12.7% on HotpotQA and 2.8%–14.9% on 2WikiMultihopQA.

Finding-8: Prioritizing knowledge pieces with rich CoE features during retrieval ensures robust reasoning even when complete evidence chains cannot be formed. Under the knowledge-complete (KC) scenario, although 27% of the questions lack data sufficient to form a fully comprehensive CoE, *ScopeCoE* preserves a high ACC of 79.7%. Comparisons with baselines substantiate *ScopeCoE*'s advantage in the KC setting: it outperforms RAG by 19.6%, VE by 51.3%, and CoT_SC by 63.6%.

Finding-9: *ScopeCoE* achieves a better trade-off between performance and computational efficiency compared to existing approaches. While implementing the CoE-guided strategy requires additional LLM calls for feature identification and discrimination, *ScopeCoE* maintains moderate computational overhead with only 2 extra calls beyond base retrieval. This is significantly more efficient than CoT_SC (5 calls) and VE (requiring additional calls for both question refinement and rationale correction, which needs at least 2 LLM calls), while still achieving superior performance across different scenarios.

A comprehensive analysis of performance variations across different model architectures is provided in Appendix D.

9 Conclusion

In this paper, we introduce CoE and investigate its impact on LLMs in imperfect external knowledge. We characterize the features of CoE knowledge and propose a CoE discrimination approach to identify CoE from external knowledge. Generally, our study reveals LLMs' preference for CoE in the imperfect context. Once CoE's implicit relevance or interconnectivity is disrupted, the preference also decreases. Furthermore, we apply CoE theory to the knowledge-intensive task, finding that retrieving CoE-structured knowledge during the retrieval phase effectively improves the response accuracy of LLMs. In future work, we will explore more scenarios where CoE can be applied.

Limitations

There are two limitations to the current study. Firstly, we apply the *ScopeCoE* 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 retrieval strategy (*ScopeCoE*) 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.

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*.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *CoRR*, abs/2304.11633.

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*.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

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.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei Koh. 2024. Scaling retrieval-based language models with a trillion-token datastore. *arXiv preprint arXiv:2407.12854*.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5823–5840. Association for Computational Linguistics.

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

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

A 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 5.

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	93.0%	84.1%	81.1%	59.9%
			0.25	93.4%	84.4%	81.6%	56.5%
			0.50	93.4%	84.8%	78.5%	54.5%
			0.75	92.6%	85.6%	76.9%	54.5%
RQ2	FR	Irrelevant	0	87.9%	85.1%	69.2%	57.3%
			0.25	79.4%	66.3%	64.8%	54.8%
			0.50	67.4%	52.0%	61.4%	53.2%
			0.75	62.7%	47.0%	58.1%	49.0%
RQ3	ACC	Misinformation	0	93.0%	84.1%	81.1%	59.9%
			0.25	86.8%	74.0%	21.7%	53.1%
			0.50	83.3%	67.4%	21.2%	52.1%
			0.75	77.5%	60.0%	19.0%	47.8%

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

Regarding faithfulness (RQ2), LLMs show the highest compliance with knowledge lacking evidence nodes, followed by knowledge missing evidence relation, and then intent. This highlights the significance of relationships and intent in guiding LLM responses. However, knowledge lacking evidence nodes demonstrates weaker resistance to irrelevant external knowledge when misinformation is present, indicating that evidence nodes play a vital role in maintaining resilience against irrelevant external knowledge under misinformation scenarios.

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.

B 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 effectiveness 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 2 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.

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	FR	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%

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.

C 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 2, 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.

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

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 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).

D Impact of Model Architectures on Usability Assessment

To investigate how different model architectures affect the performance of ScopeCoE, we conduct experiments across various LLMs. The results reveal several key patterns in performance variation:

- Model Scale Effect: Our analysis demonstrates a clear correlation between model size and performance. Larger models consistently achieve better results, aligning with their enhanced reasoning capabilities. This is exemplified by GPT-4, which achieves the highest accuracy (82.6% and 88.6%) on both datasets.

- Architecture-Specific Performance: The GPT series exhibits superior performance compared to the Llama series, primarily due to

Algorithm 1: Minimal Coverage Search

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;

their enhanced ability to capture and utilize CoE features. An interesting observation is that Qwen2-32B achieves comparable performance to Llama3-70B despite having fewer parameters, suggesting that architectural design can compensate for model size in reasoning and knowledge comprehension tasks.

- Consistent Improvement Pattern: Notably, ScopeCoE demonstrates consistent performance improvements across all model architectures when compared to RAG. The improvement margins range from 10.4% to 28.7%, highlighting the *ScopeCoE*'s generalizability across different model architectures.

E The Algorithm for the Minimal Coverage Search

We show the detailed algorithm 1 for the minimal coverage search in *ScopeCoE*.

F The Algorithm for the Minimal Coverage Search

We show the overview of *ScopeCoE* in Figure 5.

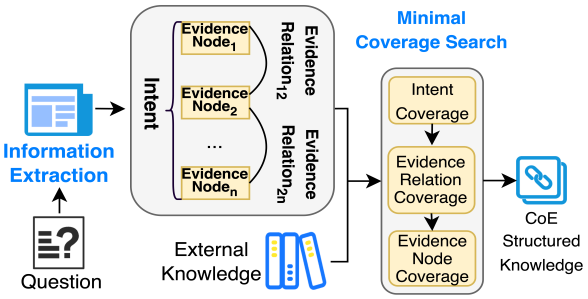

Figure 5: The overview of *ScopeCoE*.

G 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.

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:

H 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]***

I 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: