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Incorporating external knowledge into large language models (LLMs) has emerged as a promising approach to mitigate outdated knowledge and hallucination in LLMs.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 explore LLMs' preferences from their effectiveness, faithfulness and robustness, as well as CoE's usability in a naive Retrieval-Augmented Generation (RAG) case. The evaluation on five LLMs reveals that CoE enhances LLMs through more accurate generation, stronger answer faithfulness, better robustness against knowledge conflict, and improved performance in a popular RAG case.

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knowledge acquired by large language models (LLMs) through pre-training at a specific point in time becomes outdated with the knowledge evolution or produces hallucination (Achiam et al., 2023; Touvron et al., 2023a; Anil et al., 2023).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.

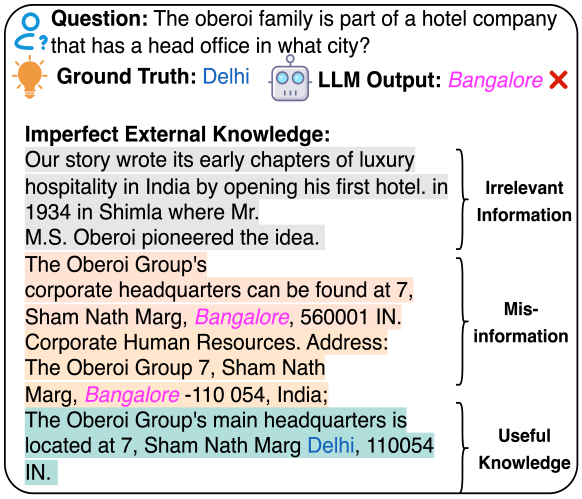

Figure 1: Example of imperfect external knowledge.

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 RAG case and design a CoE-guided retrieval strategy to explore the improvements when applying the strategy in the naive framework.

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.
- For the selected case, the CoE-guided retrieval strategy can effectively improve LLM's accuracy after substituting the reranking component in the naive RAG framework.

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 that LLMs exhibit completeness bias, manifesting in their superior performance when provided with complete external knowledge.

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.

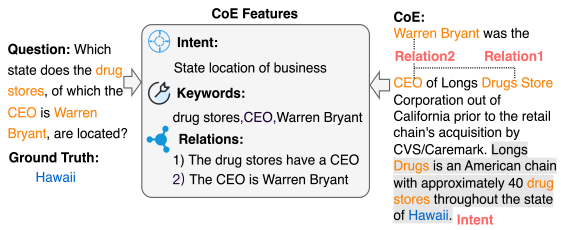

Figure 2: Example of CoE and the CoE features.

We assume that LLMs prefer knowledge that forms CoE. To satisfy the two properties required for CoE formation, we characterize three features: 1) **Intent** describes the ultimate goal the user intends to solve through the question. 2) **Keywords** are important words or phrases that capture the specific details the user is asking about; and 3) **Relations** describe how keywords are connected to each other to convey intent. Knowledge containing intent demonstrates responsiveness to the question, satisfying the relevance property, while knowledge containing keywords and relations mutually corroborates each other, fulfilling the interconnectivity property. Therefore, we consider knowledge matching the three features as CoE for the current QA, as illustrated in Figure 2.

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 3. First, for each question, we perform information extraction to extract its inherent intent, keywords, and 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 A shows the example template for the extraction prompt.

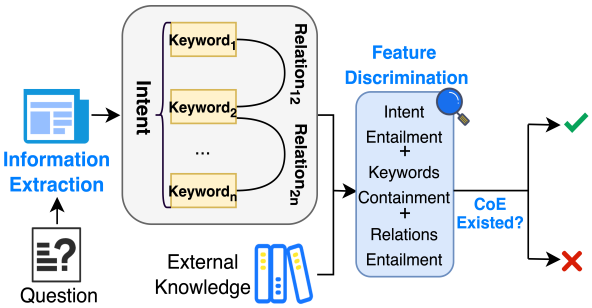

Figure 3: The overview of CoE discrimination approach.

Second, for external knowledge, the pipeline discriminates whether it contains CoE. Specifically, the approach leverages GPT-4o to discriminate the presence of intent, keywords, and 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 keywords, the LLM identifies phrases contained in external knowledge that are semantically similar with keywords. For relation entailment, the LLM utilizes its textual entailment capabilities, similar to the process of intent entailment. 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 B.

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. It is worth noting that, due to the characteristics of multi-hop QA, supporting knowledge typically contains multiple knowledge pieces[1], usually no fewer than two. Considering that supporting knowledge is initially constructed to describe the necessary information from the question to the answer, we believe it is highly likely to possess features of the CoE we have characterized. 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

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, LLMs typically require multiple knowledge pieces to generate answers. However, external knowledge is often incomplete in practice. To simulate this situation, we construct Non-CoE by removing one or more knowledge pieces from CoE. Specifically, we segment the CoE into multiple sentences and select sentences that contain keywords mentioned in the corresponding question, but not the answer, as candidates. We iteratively remove one candidate sentence at a time from the CoE and use the CoE discrimination approach to determine if the remaining knowledge doesn't contain CoE. If CoE doesn't exist in the remaining knowledge (i.e., Non-CoE), we stop the removal process; otherwise, we continue removing candidate sentences from the remaining external knowledge. Figure 4 illustrates an example of SenP.

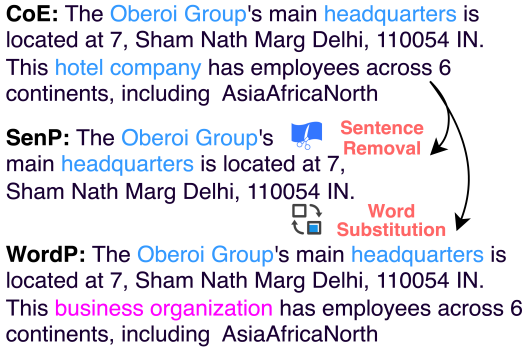

Figure 4: Examples of CoE and two types of Non-CoE.

Word-Level Perturbation (WordP). We randomly select a keyword in CoE and substitute all the mentions of the keyword with their higher-level expressions using GPT-4o, which are more generalized terms representing broader categories (for example, "hotel company" in CoE is replaced by "business organization" in Figure 4). Due to the absence of the keyword and its related relations in the perturbed CoE, it transforms into Non-CoE. WordP is a more refined substitution, which can significantly reduce the information loss of Non-CoE compared to the original CoE.

Finally, for each QA pair, we construct a five-element tuple, <*Question, Answer, CoE, SenP, WordP*>, forms 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 better help LLMs generate correct answers under external information rich with irrelevant noise.

[1] A knowledge piece refers to a complete sentence.

5.1 Experimental Setup

First, we collected the irrelevant information using the search engines. Specifically, for a constructed sample (donated as *<Question, Answer, CoE, SenP, WordP>*), we traverse all the keywords in "Question", fill them into the template "Please introduce the background of the *[keyword]*", and use Google to retrieve the knowledge snippets. In this way, we can ensure that the retrieved information is irrelevant to the question's intent and highly similar to the question in lexical terms. Then, we regard the retrieved snippets as irrelevant information and inject them into the context of "CoE", "SenP", "WordP" respectively in different ratios. Specifically, we increase the proportion of irrelevant information based on character length. We design four proportion scenarios with intervals of 0.25. Finally, "Question" together with "CoE", "SenP", "WordP" are sent to studied LLMs as input and obtain the output of the corresponding LLMs.

For each sample, we evaluate the consistency between the LLM's output and ground truth "Answer" in the five-element tuple. During the process, we followed the evaluation method used in Adlakha et al. (2024) and used GPT-4o to judge whether LLMs augmented by different external knowledge can generate the correct answer. After that, we calculated the accuracy (ACC) of each studied LLM for the three experiment groups, i.e., "CoE", "SenP" and "WordP". To alleviate the randomness of LLMs, each group of experiments is repeated three times, and the average will be taken as the final evaluation result.

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 of Non-CoE. 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).

Table 2: LLMs' Accuracy (ACC) on CoE and Non-CoE.

Model	Irrelevent 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 ($p < 0.05$)

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. In the Non-CoE, WordP demonstrates better performance over SenP, exhibiting both higher ACC and greater resistance against increasing irrelevant information. The enhanced performance of WordP, which contains richer information content than SenP, indicates that the information density of external knowledge positively correlates with LLMs' QA capabilities. Furthermore, while CoE and WordP possess comparable information content, LLMs achieve better performance with CoE, highlighting the importance of forming CoE.

In addition to the main findings illustrated above, we also observed that even under perfect retrieval conditions (Irrelevant proportion is 0%), CoE outperforms Non-CoE by 14.6% in ACC. This implies that LLMs still face challenges in utilizing external knowledge effectively, even when all retrieved information is useful.

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 in both type and format. For example, we replace "United States" with the same type "Canada", and "September 29, 1784" with the same format "April 22, 1964". We employ GPT-4o to understand the answer types and their formats, facilitating the generation of naturally incorrect answers. Appendix C presents the detailed prompt design. Through manual inspection, we found that 100.0% of the generated incorrect answers maintain the same type and format as the correct ones.

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. Each group of experiments is conducted three times and the average is considered as the final evaluation result.

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.

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 ($p < 0.05$)

Beyond the main findings, we also discovered that LLMs demonstrate a 6.6% reduction in FR when processing CoE with factual errors, compared to those with correct answers (as indicated by ACC in Table 2). This discrepancy could be attributed to the LLM's inherent parametric knowledge containing accurate information, facilitating self-correction of certain factual errors.

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

Based on the CoE and Non-CoE samples, we first obtain misinformation. Misinformation should meet two requirements: 1) contain factual errors, and 2) cause conflicts with the knowledge in CoE and Non-CoE. Following previous studies (Chen et al.; Zhou et al., 2023; Jin et al., 2024), we use two strategies to generate misinformation: 1) entity replacement, which replaces the correct answer in the CoE with the incorrect answer and uses the sentence containing this incorrect answer as misinformation; 2) LLM generation, which uses

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 ($p < 0.05$)

GPT-4o to generate multiple expressions containing the incorrect answer. Mixed with the two types of methods, we obtain all the misinformation.

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. Similarly, each group of experiments is repeated three times, and the average will be taken as the final evaluation result.

7.2 Results and Findings

Table 4 shows LLMs' response accuracy (ACC) after adding misinformation to CoE and two types 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 also discovered that as misinformation increases, LLMs with weaker reasoning capabilities tend to favor frequently appearing knowledge in external knowledge, while LLMs with stronger reasoning abilities adhere more to knowledge from CoE. With the proportion of misinformation increasing from 0% to 75%, less capable LLMs like GPT-3.5 and LLama2-13B are more likely to be misled by increasing misinformation, leading them to select answers from misinformation and resulting in significant ACC drops (with average ACC decreasing by 34.5%), whereas more powerful LLMs such as GPT-4, Llama3-70B, and Qwen2.5-32B consistently adhere to answers within CoE, resulting in slight ACC decreases (with average ACC decreasing by 7.1%).

8 Usability Assessment

To assess usability, we selected a popular knowledge-augmentation case, naive RAG, and designed a CoE-guided retrieval strategy to investigate the extent to which CoE improves the performance compared with the naive case.

8.1 Subject Case

Considering popularity and maturity, we choose a naive RAG scenario proposed by Chen et al. (2024) as our subject case. For a given question, a search engine first retrieves relevant knowledge snippets, followed by a reranking model that prioritizes the knowledge snippets based on its relevance to the question. Finally, the top K knowledge snippets are selected as external knowledge and fed into the studied LLMs to generate answers to the questions.

8.2 CoE-guided Retrieval Strategy

We design a retrieval strategy (*ScopeCoE*) guided by CoE. Instead of using the reranking component in the naive framework, *ScopeCoE* selects the minimal set of knowledge snippets that encompass a CoE as a context input for LLMs. 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.

8.2.1 CoE Feature Judgment

ScopeCoE first extracts CoE features from the question and then judges them in each knowledge snippet. Specifically, as shown in Figure 3, *ScopeCoE* employs the same information extraction component in the discrimination approach to extract the intent, keywords and 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, keywords, and relations for each knowledge snippet.

8.2.2 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 D. First, *ScopeCoE* searches for knowledge snippets that contain intent and adds them to the minimal set. Second, *ScopeCoE* examines the coverage of the relations. Specifically, it determines whether the minimal set already contains all relations. If there are uncovered relations, it searches the remaining knowledge snippets and adds those containing uncovered relations to the minimal set. Finally, *ScopeCoE* proceeds to examine keywords coverage following the same process. It checks if the minimal set covers all keywords. If uncovered keywords exist, it searches the remaining snippets for those containing these keywords.

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.

8.3 Experimental Setup

We used the constructed CoE samples (including "Question", "Answer" and "CoE") for usability evaluation. To obtain the external corpus for retrieval, we first use the Google Search API to retrieve relevant knowledge snippets for each "Question". To ensure that the corpus contains the correct answers, we decompose CoE into multiple knowledge pieces based on sentence completeness and then append them to the corpus. Then we set up two experimental groups: *RAG* and *RAG+ScopeCoE*. For *RAG*, we reuse its process in Chen et al. (2024) that selects top-5 most relevant snippets from the external corpus for LLMs' generation. For *RAG+ScopeCoE*, *ScopeCoE* is used to replace the reranking component and other processes remain consistent with *RAG*. ACC is used as the metric for assessment.

Table 5: LLMs' Accuracy (ACC) on naive RAG and RAG+*ScopeCoE*.

Model	HotpotQA		2WikiMultihopQA	
	RAG	*RAG+ScopeCoE*	*RAG*	*RAG+ScopeCoE*
GPT-3.5	68.1%	**76.0%**	54.6%	**81.5%**
GPT-4	72.9%	**82.6%**	59.3%	**88.6%**
Llama2-13B	64.4%	**74.1%**	51.7%	**74.0%**
Llama3-70B	67.8%	**79.5%**	49.4%	**80.0%**
Qwen2.5-32B	63.8%	**77.0%**	49.4%	**83.8%**

8.4 Results and Findings

Finding-7: For the subject case, CoE-guided retrieval could improve the LLMs' accuracy in the naive framework. Table 5 demonstrates the impact of naive RAG and RAG+*ScopeCoE* on LLMs' accuracy. The results show that RAG+*ScopeCoE* achieves average ACC of 77.8% and 81.6% on HotpotQA and 2WikiMultihopQA respectively, outperforming RAG by 10.4% and 28.7%.

Moreover, we also observe that *ScopeCoE* can help LLMs generate more accurate outputs with fewer knowledge pieces (4.6 for HotpotQA and 4.8 for 2WikiMultihopQA) compared to the naive framework (5 pieces). It implies that *ScopeCoE* can make LLMs more efficient in knowledge utilization, leading to improved performance and reduced dependency on large amounts of external data.

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 naive RAG framework, finding that retrieving CoE-structured knowledge during the retrieval phase effectively improves the response accuracy of LLMs. In future work, we will explore broader applications of CoE in RAG scenarios, such as retrieval corpus construction and retriever optimization.

Limitations

There are three 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 the 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, this paper does not investigate the individual contributions of CoE features to LLM performance. Since intent, keywords, and relations within CoE are interdependent, it is challenging to isolate any single feature. Therefore, we focus on examining the overall impact of CoE on LLM performance in this paper.

Thirdly, 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.

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*.

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

A 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 keywords.

Intent and Keyword Extraction Prompt:
Please extract both the intent and keywords 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 keywords, 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", "Keywords": ["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", "Keywords": ["Oberoi family", "head office"] }
Example3:
Question: What nationality was James Henry Miller's wife?
Output: { "Intent": "Nationality of person", "Keywords": ["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", "Keywords": ["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", "Keywords": ["Malcolm Smith", "Most Valuable player"] }
Question: ***[Question]***
Output:

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

B 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, keyword, and 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]***

Keyword Discrimination Prompt:
Please determine if the input keyword 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 Keyword: ***[Keyword]***
Input external knowledge: ***[External Knowledge]***

Relation Discrimination Prompt:
Please infer whether the input external knowledge can infer the input relation description. If there is definite evidence in the input sentence to prove that the input relation description is true, then output "yes", otherwise output "no". Please output only "yes" or "no".
Input relation description: ***[Relation]***
Input external knowledge: ***[External Knowledge]***

C 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:

Algorithm 1: Minimal Coverage Search

Input: External knowledge list EK, Judged external knowledge list IEK, where each item contains Intent, Relations, and Keywords judgments
Output: Set S of minimal coverage external knowledge

```
S ← ∅;
# Phase 1: Intent Coverage;
for i ← 0 to |IEK| − 1 do
    if IE[i].Intent = TRUE then
        S ← S ∪ {EK[i]}
# Phase 2: Relation Coverage;
R_uncovered ← GetUncoveredRelations(IEK, S);
for r ∈ R_uncovered do
    for i ← 0 to |IEK| − 1 do
        if IEK[i].Relations[r] = TRUE then
            S ← S ∪ {EK[i]};
            break;
# Phase 3: Keyword Coverage;
K_uncovered ← GetUncoveredKeywords(IEK, S);
for k ∈ K_uncovered do
    for i ← 0 to |IEK| − 1 do
        if IEK[i].Keywords[k] = TRUE then
            S ← S ∪ {EK[i]};
            break;
return S;
```

D The Algorithm for the Minimal Coverage Search

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