Chain-of-Discussion: A Multi-Model Framework for Complex Evidence-Based Question Answering

Anonymous ACL submission

Abstract

Open-ended question answering requires models to find appropriate evidence to form well-reasoned, comprehensive and helpful answers. In practical applications, models also need to engage in extended discussions on potential scenarios closely relevant to the question. With augmentation of retrieval module, open-source Large Language Models (LLMs) can produce coherent answers often with different focuses, but are still sub-optimal in terms of reliable evidence selection and in-depth question analysis. In this paper, we propose a novel Chain-of-Discussion framework to leverage the synergy among multiple open-source LLMs aiming to provide **more correct** and **more comprehensive** answers for open-ended QA, although they are not strong enough individually. Our experiments show that discussions among multiple LLMs play a vital role in enhancing the quality of answers. We will release our data and code for further research.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable language generation capabilities (Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023), propelling advancements in various understanding/generation tasks, including open-domain question answering (QA) (Song et al., 2024). However, for complex open-ended question answering, which plays an important role in human-AI interaction, LLMs may still produce output with hallucination and often deliver inferior performance compared to short-form QA (Huang et al., 2023a). This task usually requires LLMs to analyze the questions first, retrieve evidence accordingly, then form a long-form answer which is expected to be correct and well-reasoned with details and proper evidence supported. It has a wide range of applications, from legal consultations and medical advice to education support and financial

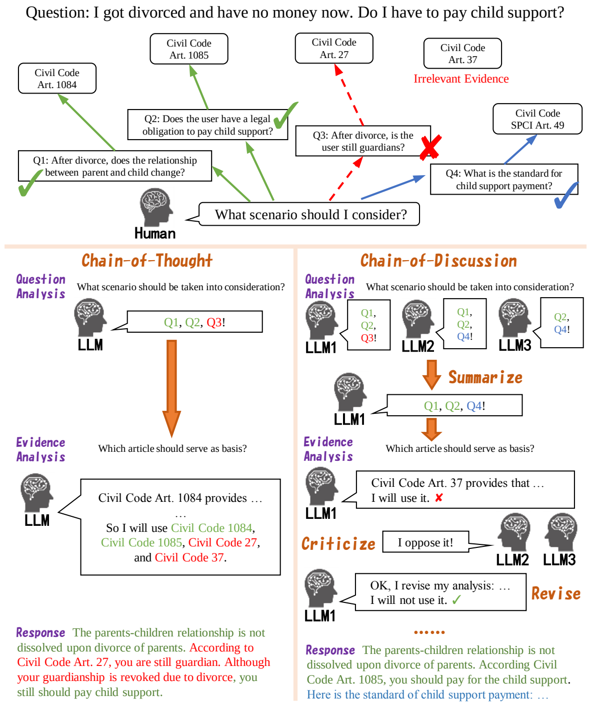

Figure 1: The process of Chain-of-Discussion, compared with chain-of-thought. The green parts are necessary to answer user's question. Blue parts indicate closely related to the question, useful for detailed/extended discussions. Red parts are irrelevant content that should be avoided.

analysis, where users may pose various complex and knowledge-intensive questions.

Although current LLMs can produce long and coherent texts (Peng et al., 2024), the complex open-ended QA is still an admittedly challenging task, even with augmented retrieval modules. The challenges primarily arise from two aspects.

Firstly, retrieval models are not entirely prefect, inevitably with noise in the retrieval results. As a legal consultation example shown in Figure 1, the model is required to respond to a question regarding the necessity of child support payments. Due to the semantic similarity between obligations for supporting children (financially) and raising/protecting children (physically), the retrieval model may wrongly return law articles pertaining to guardianship qualifications. LLMs usually cannot filter all these noisy evidence, which may propagate and lead to incomplete analysis, wrong reasoning paths, biased opinions and finally problematic or even misleading answers.

Secondly, we expect LLMs output correct responses and consistent explanations, providing useful suggestions about potential scenarios not directly mentioned in questions but indeed helpful for users' current/near-future situations. In Figure 1, when responding to a question about the obligation to pay child support for a user facing financial difficulties, the model should also remind her/him of the standards for child support payments and ways to negotiate for a reduction given her/his current situation. This is even hard for humans where one should have access to proper evidence, e.g., necessary or closely related law articles, and accordingly provide kind reminders with reasonable explanations. Let alone LLMs without specific training/fine-tuning, which usually focus on the specific facts literally appearing in the questions.

In this work, we focus on the complex evidence-based question answering (CEBQA) task, a typical example of the open-ended QA tasks. We collect a high quality CEBQA dataset consisting of 200 carefully annotated legal consultation questions in the field of marriage and family affairs. To address the challenges, we propose a novel chain-of-thought (CoT) framework, Chain-of-Discussion (CoD), encouraging multiple LLMs to summarize, criticize and revise each other's output to reach a well-supported and helpful response.

Our motivations are two-fold. First, different LLMs may have different intrinsic knowledge and reasoning capabilities due to different training data. Thus, multiple LLMs can be less possible to make errors concurrently than a single LLM. Recent works (Zhang et al., 2023) show checking the consistency across multiple LLMs helps reduce output hallucinations. Specifically, we propose a criticize-and-revise framework, which requires multiple LLMs to discuss and reach a consensus for a better response. For questions that need to involve helpful scenarios or possible extensions, we guess multiple LLMs may provide a diverse set of perspectives to address these possibilities. We thus propose a summarizing step to gather different but helpful perspectives from multiple LLMs, which will eventually form comprehensive and detailed responses based on the summarized analyses.

Different from existing multi-model interaction works (Chan et al., 2024; Zhang et al., 2023) using strong closed-source LLMs, e.g., GPT-4 (OpenAI, 2023), we decide to take a challenge to study how to best exploit the small-scaled open-source LLMs, e.g., around 7B parameters, for a shared objective, while pushing the boundary of research regarding multi-model interaction.

Our main contributions are as follows: (1) We collect a high-quality CEBQA dataset consisting of 200 legal consultation questions in Chinese with carefully annotated evidence and answers. (2) We propose a novel chain-of-discussion framework, i.e., summarize-criticize-revise, which harnesses the synergy among multiple open-source LLMs to generate more accurate and helpful responses. (3) Both GPT-4-based and evidence-centric evaluations demonstrate our framework can help small-scaled LLMs benefit from each other and improve the overall quality in terms of correctness and comprehensiveness.

2 Related Works

Retrieval-Augmented Generation Lewis et al. (2020) initially propose the paradigm of retrieval-augmented generation (RAG), which can effectively reduce hallucinations within the texts generated by LLMs. RAG offers a vital solution to mitigate the problem of LLMs lacking domain-specific knowledge, thereby enhancing the credibility of LLMs (Gao et al., 2023). In the RAG paradigm, models typically undergo multiple generation steps to achieve the final results. For a user input, models first run a retriever to scan the store of evidence to select several documents as reference. Subsequently, models should determine when and whether to use each evidence document before generating (Izacard et al., 2022; Shi et al., 2023b; Yu et al., 2023; Trivedi et al., 2023).

In this work, we face more complex challenges than RAG. While the model filters out irrelevant evidence, it also needs to retain evidence relevant to potential scenarios. Sometimes, determining which evidence can be used for potential scenarios and which are irrelevant is also challenging for humans.

Chain-of-Thought Previous works demonstrate that LLMs have a promising capability to decompose a complex question into several intermediate steps (Wei et al., 2022; Kojima et al., 2022). By segmenting the original question, LLMs can focus on handling each simple sub-question at each step, thus yield more accurate results (Zhou et al., 2023). The CoT framework is now widely employed in diverse practical NLP applications (Zelikman et al., 2022; Shi et al., 2023a; Wang et al., 2023). Previous works also employ CoT in the self-correction process of LLMs, which aims to re-generate better outputs. For instance, in Chain-of-Verification, the model generates several queries to verify its original answer, and then revise the answer based on the verification results (Dhuliawala et al., 2023). Most of these efforts perform self-checking based on a single model. However, we study a novel CoT framework for multi-model interactive checking and re-generating.

3 Preliminaries

Task Definition In CEBQA tasks, given a user's question q and a store of evidence documents $\mathcal{D}$, a model should analyze q first, find necessary evidence $\mathcal{D}_q = \{d_1, \cdots, d_t\}$ from $\mathcal{D}$ accordingly and generate a paragraph r as the final response. For instance, in the legal consultation task, users may ask what to do given her/his current situation. The model should find supportive evidence from a store of law articles or previous legal cases, and generate a helpful and detailed response.

Specifically, we expect the generated responses to meet the requirements in terms of correctness and comprehensiveness. (1) **Correctness:** The responses should be based on the evidence that can support to answer the questions, and refrain from employing irrelevant evidence or misinterpreting evidence out of context. (2) **Comprehensiveness:** The responses should engage in discussions about potential scenarios that would be relevant/helpful to users, even not explicitly mentioned in questions.

We note that it is hard to guarantee all the retrieved evidence pieces can be perfectly used to answer the question. Therefore, similar to RAG, models should filter out irrelevant evidence. However, it is more challenging for models to carefully retain the evidence that can be used for discussions about potential scenarios, even though the evidence may not directly support answering the question.

Baseline Framework: CoT Previous works have revealed that the CoT prompt can enhance the ability of LLMs to handle complex reasoning tasks (Wei et al., 2022; Kojima et al., 2022). Inspired by these works, we employ a multi-step prompt to stimulate LLMs to generate more correct while comprehensive answers.

We initially prompt LLMs to analyze the question q, including identifying the possible role of users, understanding explicit and implicit demands of users, and determining what types of evidence is needed to answer the question. The generated analysis of question can be denoted as a_q^{que}.

The next step is to judge whether each evidence document can serve as a potential basis for responding to the question q. Here, we employ a prompt to feed the LLM with question q, analysis a_q^{que} of the question, and a specific evidence document d_i. The LLM then need to analyze whether $a_{d_i}^{\text{evi}}$ can be used to address the issues raised in q and whether evidence d_i can probably be used to respond or not.

The LLM with parameters θ should finally respond to the question q according to question analysis a_q^{que} and evidence analysis $\left\{a_{d_i}^{\text{evi}}\right\}_i$, based on the evidence document set $\mathcal{D}_q$:

$$r = f\left(q, \mathcal{D}_q, a_q^{\text{que}}, \{a_{d_1}^{\text{evi}}, \cdots, a_{d_t}^{\text{evi}}\} \,|\theta\right).$$

As observed in our pilot study, one small-scaled LLM could generate fluent answers, but often with incomplete analysis or wrong reasoning paths.

4 CoD: Summarize, Criticize, and Revise

Our Chain-of-Discussion framework leverages interactive discussions among multiple LLMs, thereby addressing potential shortcomings in individual's intrinsic knowledge.

Similar to the baseline, we employ a two-stage analyzing pipeline that instructs LLMs to analyze the question and evidence separately. To address the correctness and comprehensiveness of generated answers, during question analysis, we encourage models to read and summarize others' analyses so as to consider more scenarios closely relevant to the question, in the purpose of augmenting the comprehensiveness. During the stage of evidence analysis, we require all other LLMs to **criticize** the evidence analysis of each LLM. Subsequently, the model will read others' critique and determine whether to **revise** its own analysis or not. The model finally generate a correct and more helpful response based on the summarized question analysis and revised evidence analysis.

4.1 Stage 1: Question Analysis

Formally, suppose there are n accessible LLMs, denoted as $M_1, \cdots, M_n$. For a given question q and the retrieved evidence $\mathcal{D}_q$, we aim to employ the target LLM M_k to generate a response, with the assistance of the remaining LLMs.

We first instruct the LLMs to analyze the question, including facts mentions in q, primary needs of the user, and potential scenarios associated with the question. We observe that LLMs may perform poorly in analyzing potential scenarios when solely relying on their intrinsic knowledge, especially those models that have not been pre-trained or supervised fine-tuned on domain-specific data. Thus, we argue that the evidence documents $\mathcal{D}_q$ can serve as vital cues about the potential scenarios not mentioned in q.

Different LLMs can have varying preferences in analyzing the potential scenarios. Therefore, we believe that by integrating the outputs of multiple LLMs, we can take more helpful scenarios into account, thus improve the **comprehensiveness** of question analysis. We prompt each LLM M_i to analyze the question q, with retrieved evidence $\mathcal{D}_q$ as a reference: $a^{\text{que}}_{q,\,M_i} = f_{\text{que}}\left(q,\ \mathcal{D}_q|\theta_{M_i}\right)$.

We then employ the target LLM M_k to **summarize** the question analyses of all models, according to following instructions:

- **Consistency:** If the majority of LLMs provide similar analyses regarding a fact in the question or a potential scenario, then it is likely to be correct. You can include it in the summary.
- **Comprehensiveness:** If a minority of LLMs hold a particular viewpoint in their analyses with reasons, it does not imply its unreliability. You should scrutinize this content, assessing its logical coherence and relevance to the question.

The summarized question analysis can be $a^{\text{que}}_q = f_{\text{sum}}\left(q,\ a^{\text{que}}_{q,\,M_1},\ \cdots,\ a^{\text{que}}_{q,\,M_n}|\theta_{M_k}\right)$.

4.2 Stage 2: Evidence Analysis

Incorporating many irrelevant evidence documents as input would inevitably introduce noise, which could deteriorate the model performance. Thus, we should discern which evidence document should be used to address the question. For an evidence document $d_j \in \mathcal{D}_q$, we prompt the target model M_k to analyze it based on the question and question analysis : $\hat{a}^{\text{evi}}_{d_j} = f_{\text{evi}}(d_j,\ q,\ a^{\text{que}}_q|\theta_{M_k})$.

However, a single LLM might generate hallucinated outputs (Li et al., 2023b; Huang et al., 2023a), and incorrectly assess the relevance between evidence documents and the given question. Inspired by previous work (Zhang et al., 2023), we propose a multi-party discussion framework to improve the quality of evidence analysis.

First, we instruct each LLM, excluding M_k, to **criticize** the evidence analysis $\hat{a}^{\text{evi}}_{d_j}$. Each critic model M_i should explicitly output whether it holds opinions contrary to $\hat{a}^{\text{evi}}_{d_j}$, which are denoted as $c^{d_j}_i$. In this work, we employ a revising threshold δ. If the proportion of opposite opinions in the critiques exceeds δ, the target model needs to **revise** its evidence analysis: $a^{\text{rev}}_{d_j} = f_{\text{rev}}\left(q,\ d_j,\ a^{\text{que}}_q,\ \hat{a}^{\text{evi}}_{d_j}|\{c^{d_j}_i\}_i,\ \theta_{M_k}\right)$.

We assume that the critique requiring to revise can be reliable only when a majority of critic models achieve a consensus. Otherwise, we retain the original evidence analysis. Formally, we collect the evidence analysis as following:

$$a^{\text{evi}}_{d_j} = \begin{cases} \hat{a}^{\text{evi}}_{d_j}, & \text{if } \frac{|\{c_i|c_i=\text{opposite}\}|}{|\{c_i\}|} \leq \delta; \\ a^{\text{rev}}_{d_j}, & \text{otherwise.} \end{cases}$$

4.3 Response Generation

For a fair comparison, we employ prompts similar to those of the baseline framework to generate responses. We denote the response as $r = f_{\text{ans}}\left(q, \mathcal{D}_q, a^{\text{que}}_q, \{a^{\text{evi}}_{d_1}, \cdots, a^{\text{evi}}_{d_t}\}\,|\theta_{M_k}\right)$.

5 Experiments

As there is no existing dataset for CEBQA tasks, here we delve into the legal consultation task and collect the first CEBQA dataset. In civil law systems like China, all legal activities, including legal consultation, should be based on *law articles* (or *judicial interpretations*), which can be naturally considered as the evidence store in our framework.

5.1 Data Collection

To reflect the diversity in practical scenarios, we focus on the fields of *marriage, family affairs, and inheritance*, covering various legal disputes, e.g., *divorce, custody, contracts, property*, etc. From a pool of 609 questions collected from real users with consultants' responses through Web Search Engines, we employ an legal expert to select 200 questions to ensure their semantic distinctiveness and coverage. See more details in Ethics Statement.

Evidence Annotation We construct the evidence store based on all 1,749 articles of the *Civil Code*, *Civil Procedure Law* and their judicial interpretations in China. For each question, we consider three types of articles: *necessary*, *optional*, and *not required*. *Necessary* articles are highly relevant to the question, while *optional* ones can be basis for discussions of potential scenarios (more details in Appendix B). We ensure there are 5 articles retrieved for each question, and on average, each contains 1.52 *necessary*, 1.23 *optional*, and 2.25 *not required* articles. It means about 45% of the retrieved articles are not required at all.

Data Quality We employ 6 annotators with background in civil law to manually check the questions, answers and articles. Annotators are instructed to correct all typos but retain the informal expressions in questions. Note that there are many omissions or slight word-order inversions in the questions, posing a challenge to models' reasoning capabilities.

Annotators also examine the correctness and logical coherence of the responses. For problematic ones, annotators are encouraged to discuss and reach a consensus for modifications, otherwise, leave them as they are. Averagely, it takes about 20 hours per annotator to examine 100 instances.

5.2 Experimental Setup

As our main focus is to investigate whether small open-source LLMs can collaborate through summarize-criticize-revise, we study four open-source fine-tuned LLMs, Baichuan2-7B (Baichuan, 2023), Deepseek-7B (DeepSeek-AI, 2024), Qwen-7B (Bai et al., 2023), and Xverse-7B[1], which are four of the best 7B-parameter LLMs performing on CMMLU (Li et al., 2023a). When we use a specific LLM as the target model, the other three LLMs are expected to generate diverse question analyses and criticize the evidence analysis of target model.

To examine if close-source LLMs can still benefit from our CoD, we also test with *gpt-3.5-turbo-1106*, *gemini-1.0-pro-latest*, and *claude-3-haiku-20240307* similarly to the open-source group.

We note that the two stages in Chain-of-Discussion are independent of each other. Therefore, we can investigate how they contribute to the ultimate performance by the following settings:
Single-model baselines (BS): One LLM takes a query and several retrieved articles as input and performs question analysis, article analysis, and response generation in a vanilla CoT manner.
Only Stage 1 (S1): All LLMs produce question analysis. The target LLM summarizes these analyses, and proceeds to the rest by itself.
Only Stage 2 (S2): Three other LLMs criticize the article analysis generated by the target LLM. The question analysis and the final response are generated by the target LLM on its own.
Chain-of-Discussion (S1S2): All LLMs involve into both question analysis and article analysis. Eventually, the target LLM produces the response by itself.

We employ each LLM as the target model, replicating the experimental settings. We report the performance for each LLM as the target role. More details and prompt templates are in Appendix A and E.

Evaluation Metrics Different from conventional QA tasks, the responses in CEBQA tasks can consist of several hundred or even thousands of words, which are also knowledge intensive and complex in structures, containing facts and causal relations to be verified. Therefore, it is impossible to employ the popular metrics such as F1 or exact match (Joshi et al., 2017; Rajpurkar et al., 2018).

Following previous works (Liu et al., 2023; Chan et al., 2024), we employ GPT-4 to evaluate the quality of generated responses, with the expert-written responses, necessary and optional articles as reference. We prompt `gpt4-turbo-0125` to score a response in an integer between 1 and 10 based on correctness and comprehensiveness. If there is no clear reason to indicate a response is significantly better or worse than human-written ones, a score of around 7 should be given (Scoring prompt is in Appendix F).

5.3 Main Results

Table 1 shows the evaluation results produced by GPT-4. Comparing the baseline CoT setting (BS) and Chain-of-Discussion (S1S2) in the open-source group, we can find **each LLM can obtain improvements from discussions with other LLMs**, with Baichuan2-7B increased by +0.340, Deepseek-7B by +0.115, Qwen-7B by +0.120, Xverse-7B by + 0.110. We think other LLMs bring more related aspects according to their own strengths into discussions while the baseline COT setting has to rely on one LLM only. We also find employing multi-model discussion on both stages can bring more improvement than using it on one stage only.

Although CoD can enhance all LLMs, the CoD-augmented Baichuan2-7B, Qwen-7B, or Xverse-7B still can not outperform Deepseek-7B under its baseline setting, with around 0.5 left behind. This

[1] `https://huggingface.co/xverse/XVERSE-7B-Chat`

Target LLM	Setting	Avg. Score	ΔScore
Open-source LLMs Group			
Baichuan2-7B	BS	5.750	–
	S1	6.030	+0.280
	S2	5.935	+0.185
	S1S2	**6.090**	+0.340
Deepseek-7B	BS	6.465	–
	S1	6.505	+0.040
	S2	6.480	+0.015
	S1S2	**6.580**	+0.115
Qwen-7B	BS	5.835	–
	S1	5.890	+0.055
	S2	5.815	- 0.020
	S1S2	**5.955**	+0.120
Xverse-7B	BS	6.015	–
	S1	5.995	- 0.020
	S2	6.030	+0.015
	S1S2	**6.125**	+0.110
Close-source LLMs Group			
GPT-3.5-turbo	BS	6.895	–
	S1S2	**6.955**	+0.060
Gemini-1.0-pro	BS	6.940	–
	S1S2	**6.975**	+0.035
Claude-3-haiku	BS	7.280	–
	S1S2	**7.300**	+0.020

Table 1: The average scores of each target LLM and each setting evaluated by GPT-4. The upper is evaluated within the open-source LLMs group, while the bottom is within the closed-source LLMs group.

Target LLM	Setting	`N-Acc%`	`O-Acc%`
Baichuan2-7B	BS	58.26	50.14
	S1	60.03	50.67
	S2	61.86	50.25
	S1S2	**63.17**	**52.38**
Deepseek-7B	BS	75.93	59.27
	S1	76.36	59.70
	S2	76.12	59.23
	S1S2	**76.79**	**59.80**
Qwen-7B	BS	69.87	60.98
	S1	70.31	61.63
	S2	70.64	63.65
	S1S2	**71.29**	**64.20**
Xverse-7B	BS	74.00	63.95
	S1	74.24	64.72
	S2	75.67	64.44
	S1S2	**76.16**	**65.35**

Table 2: The Macro average `N-Acc` and `O-Acc` results of each target LLM and each setting. The highest scores are made **bold**, while the second underlined.

may indicate that the quality of responses primarily relies on the inherent ability of LLMs to analyze context and then to generate.

As shown in the bottom of Table 1, even for powerful closed-source LLMs like Claude-3, our CoD framework can still brings improvement consistently, but the improvements are admittedly small compared to the open-source group.

Through a manual check, we also find that these powerful closed-source LLMs may also produce erroneous analyses of questions, resulting in incorrect responses. We believe that regardless the model scales, our CoD framework can leverage the varied knowledge and capabilities of multiple models to achieve better performance.

6 Discussions

6.1 Evidence-Centric Evaluation

Besides overall evaluation by GPT-4, we wonder if our Chain-of-Discussion framework can enhance the comprehensiveness and correctness of the model output. When discussing the details of questions or potential scenarios, LLMs should refer to *necessary* or *optional* evidence. Hence, we can assess the correctness and comprehensiveness of responses by the accuracy of evidence documents.

We accordingly design two metrics of accuracy, `N-Acc` and `O-Acc`, to assess the correctness and comprehensiveness, respectively. We utilize the *not required* articles as negative samples. For `N-Acc`, we employ the *necessary* articles as positive samples, while the *optional* articles for `O-Acc`. We employ heuristic methods to examine if a response has used an article (details in Appendix C).

We compute the Macro average `N-Acc` and `O-Acc` across all examples, shown in Table 2. We can see these LLMs exhibit differently on their own, and when performing a simple majority vote among all LLMs, we could get 77.62% for `N-Acc`, confirming our hypothesis that **it is possible to harness the synergy among multiple open-source LLMs for a better performance**. Compared to the baselines (BS), our CoD (S1S2) can achieve around a 2% improvement on both `N-Acc` and `O-Acc` for Baichuan2-7B, Qwen-7B, and Xverse-7B. Even for Deepseek-7B, the best in GPT-4 based evaluation, our framework still brings improvements of 0.86% and 0.53% to `N-Acc` and `O-Acc`, respectively.

The results indicate that introducing multi-model discussions during both question analysis and evidence analysis **helps target LLMs better refer to correct evidence**. This also explains why CoD can improve the quality of model responses.

Comparing the results under BS, S1, and S2, we find that involving multiple LLMs in a single stage can actually enhance both correctness and comprehensiveness. Specifically, employing multi-model

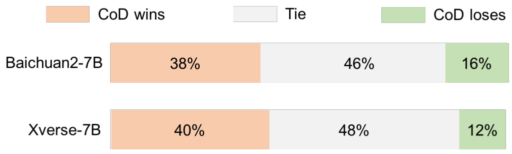

Figure 2: Human preference evaluation, comparing the CoD settings of Baichuan2-7B and Xverse-7B to their corresponding baseline settings across 50 randomly sampled examples.

discussions in question analysis contributes more to comprehensiveness, while introducing other models in evidence analysis (BS vs. S2) helps filtering irrelevant evidence thus brings more improvement in correctness.

6.2 Human Evaluation

We further examine the quality of generated responses through a win-rate analysis by human experts. We randomly sample 50 examples, and ask two human experts to examine whether the responses of CoD-enhanced models are better than those of baselines. We use the responses of Baichuan2-7B and Xverse-7B, which are models obtaining the most and the least improvement from CoD, respectively.

As shown in Figure 2, in 38% of the examples, CoD can bring improvements to the responses of Baichuan2-7B, while 40% responses generated by Xverse+CoD surpass Xverse's baseline setting. For Baichuan2-7B, only 16% of responses deteriorated after introducing the CoD framework, while for Xverse, only 12% worsened.

Qualitative Analysis. We also manually examine 30 responses of both models. The main error types are *misunderstanding key legal terms* (60%), *discussing not-required articles* (15%), and *inconsistent explanations* (10%). The latter is often caused by term misunderstanding. This shows that there may not be sufficient legal knowledge in 7B models to distinguish very similar terms. Comparing our CoD and baseline setting, we can see CoD can correct improper articles for 6 out 30 cases, leading to better results, through the cooperation mechanism.

Correlation between GPT-4 Evaluation and Human Although GPT-4 has been deemed as a reliable evaluator in various LLM evaluations (Liu et al., 2023), we still want to examine how well GPT-4 preferences correlate to human experts in our challenging task. For each example from the randomly sampled 50 cases: (1) if GPT-4 assigns a higher score for the response generated by the CoD setting, we denote this evaluation result as +1; (2) if GPT-4 assigns a higher score for the response by the baseline setting, we denote this evaluation result as -1; Otherwise, we denote it as 0. We use the same principle to annotate the evaluation results of human expert.

The Pearson correlation coefficient between GPT-4 scores and human expert evaluation is 0.5863 with $p = 7.7 \times 10^{-6}$. This indicates a relatively high correlation between the two evaluation methods, where we could consider GPT-4 evaluation as a reliable manner for our work.

6.3 Limitations of Open-Source LLMs

As shown in Table 1, using multi-model interaction at only one stage fails to enhance Xverse-7B (S1) or Qwen-7B (S2). We guess this may be due to the small size of the model scales, 7B here.

After a manual check, we find the main difficulties for those small scaled open-source LLMs: 1) those 7B LLMs may not have sufficient legal/background knowledge and reasoning capabilities, thus often fail to discern errors in the analysis of others and integrate them into the summary; 2) the small scaled LLMs do not always follow instructions, and often refuse to modify the errors in evidence analysis.

Hallucination Propagation in Summary. Table 3 illustrates an example where Baichuan-7B incorporates inaccurate analysis from DeepSeek-7B into the summary.We find that both LLMs can correctly identify the primary intention from the question, i.e., determining whether providing support for parents is a prerequisite for inheriting. Unfortunately, Deepseek-7B mistakenly engages in analysis regarding whether parents pay support for children, while Baichuan2-7B fails to identify this as irrelevant to the question.

Although both raising children and supporting parents are financial support and life care between two generations in a family, it is a commonsense, for people with legal background, that these are two different concepts and legally, there is no causal relationship between them. We think both LLMs actually rely on shallow semantic similarity to judge if a question analysis is relevant to user's question.

Capturing Nuance between Related Terms. We guess that open-source LLMs with around 7B parameters may not have sufficient knowledge and enough reasoning capabilities to distinguish related but different terms. Table 4 shows an example where Qwen-7B fails to recognize that 复婚 (rec-

Question: Can a daughter/son still be eligible to inherit if she/he has not supported her/his parents?

Original question analysis:
Baichuan2: This question involves the relationship between inheritance rights and filial duties. The user may be concerned about whether a person who has not fulfilled filial duties still qualifies for inheritance. This may entail rules regarding the recognition of inheritance rights, the forfeiture and reinstatement of inheritance rights, and the role of filial duties in the determination of inheritance rights.

Question analyses of other LLMs:
Qwen: ... **Xverse:** ...
Deepseek: ... The user may want to clarify whether not supporting parents will affect her/his eligibility to inherit parents' estate. Additionally, user may also inquire about her/his entitlement to demand child support payments from her/his parents, if the parents refuse to raise children. ...

Summarized question analysis:
Baichuan2: This question involves the relationship between inheritance rights and filial duties. The user may wonder if her/his child retains eligibility for inheritance despite failing to fulfill filial duties. This may entail rules regarding the recognition of inheritance rights, the forfeiture and reinstatement of inheritance rights, and the role of filial duties in the determination of inheritance rights. The user may also want to know the regulations regarding whether parent who does not raise children is required to pay child support.

Table 3: An example of hallucinated summary of question analysis. Violet texts are correct analysis, while the texts with yellow background are hallucinated parts.

oncile and remarry) and 再婚 (remarry with other person) represent different marriage relationships, comply with different articles or provisions, and should not be used interchangeably. However, in our CoD setting, the discussions with other models can, to some extent, have the chance to reduce the hallucination caused by the target model's poor reasoning capabilities, thus help to reduce logical errors in Qwen-7B's responses.

Reasoning with Knowledge. Similarly, we find the lack of good reasoning capabilities and knowledge may also limit the performance of LLMs in revising evidence analysis. LLMs tend to focus on the literal differences between questions and law articles. Users often omit commonsense information in their questions, whereas law articles exhibit a rigorous and detailed structure. The stylistic differences between questions and evidence may lead to the misjudgments of LLMs (more examples and analysis in Appendix D.2).

6.3.1 Model Preference of Revising

We find different LLMs have varying preferences regarding whether to modify the evidence analysis.

To quantitatively study this issue, we initially employ Baichuan2-7B to generate the original evi-

Question: What is the difference between *resumption of marital relationship* (复婚) and *remarriage with other person* (再婚)?
Article: *Article 1046* A man and a woman shall enter into marriage freely and voluntarily. ...
Article 1083 Where, after divorce, both the man and the woman voluntarily intend to resume their marital relationship, they shall file for re-registration...

Qwen-7B: ... According to Article 1083, both parities should be voluntary for *resumption of marital relationship*, while there is no such limitation for *remarriage with other person*. ...
Qwen-7B+CoD: ... According to Article 1046, whether it is *Fuhun* or *Zaihun*, both parities need to do so voluntarily.

Table 4: A case of Qwen-7B obtaining improvement from CoD. Violet texts are correct analysis, while the texts with yellow background are hallucinated parts.

dence analyses. Deepseek-7B and Xverse-7B then play the role of critics to determine which analysis should be revised. Then, we use Baichuan2-7B and Qwen-7B to revise these analyses, respectively. We find that Baichuan2-7B successfully revises 96.5% of the analyses, while Qwen-7B can only revise 56.1% of them.

To mitigate the influence of target model selection, we also use Qwen-7B to provide origianl evidence analyses, still with Deepseek-7B and Xverse-7B as the critics. Similarly, Baichuan2-7B can revise 92.5% of the analyses, but Qwen-7B only revise 67.2% of them.

We argue that an LLM's preference for refusing to revise may lead to a failure to obtain better evidence analysis based on the critiques. Consequently, it might limit our Chain-of-Discussion framework to bringing more improvement as expected. The preference of LLMs can be affected by supervised fine-tuning and reward modeling (Ouyang et al., 2022; Rafailov et al., 2023). We hope to study the effect of supervised training on Chain-of-Discussion in future.

7 Conclusions

In this work, we proposed a novel reasoning framework, Chain-of-Discussion, for complex evidence-based question answering tasks. The CoD framework involves multiple LLMs in discussions to achieve more correct and comprehensive responses with less hallucination and more supportive evidence. Experiments on a challenging legal consultation dataset show CoD can effectively improve the performance of open-source LLMs by encouraging them to discuss and criticize.

Limitations

Our proposed framework is designed to generate correct and comprehensive answers to respond complex questions. When used for providing legal advisory services, this technique can produce helpful responses to help people with needs, but it still cannot guarantee all responses are completely correct. Hence, this techniques should be used with cautions for further applications.

Our dataset is designed and annotated to reflect the nature of CEBQA tasks, which requires models to generate detailed analysis to each closely relevant scenarios of the user's question. However, our annotated results may be inevitably not perfect from the professional perspectives of experts in civil law. Thus it should be used with caution and for research purpose only.

We also note that the proposed framework involves multiple LLMs to generate for several rounds. Straightly using commercial APIs may lead to more promising generated results and cost less time. However, our aim is to validate how to better and more efficiently exploit the synergy among small LLMs, without relying on larger LLMs. We pioneer to expand the border of investigation about multi-model interaction to the small open-source LLMs.

Ethics Statement

In this work, we collect the legal consultation question-answer pairs through Web Search Engine. All these data are publicly accessible on the internet. Additionally, these websites all allow crawlers of search engines to automatically access their web pages.

To protect the privacy of the consultants in real world, we manually check all the instances, and remove those which may indicate the name, address, phone number of a specific users. Note that information such as gender, age, and family composition serves as crucial background in judicial practice. Since this information does not identify a specific individual and does not result in privacy leaks, we do not discard data where the users mention their gender, age, or family composition.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: An open-source neural language technology platform for Chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023b. Lawyer llama technical report.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2023. Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023a. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023b. Replug: Retrieval-augmented black-box language models.

Yongho Song, Dahyun Lee, Myungha Jang, Seung won Hwang, Kyungjae Lee, Dongha Lee, and Jinyeong Yeo. 2024. Evidentiality-aware retrieval for overcoming abstractiveness in open-domain question answering. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023. Boosting language models reasoning with chain-of-knowledge prompting.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STar: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Details of Generation

During generation, we set the temperature to 0.3, the repetition penalty to 1.05, and the top-p to 0.8. At Stage 2, the revising threshold δ is set to 0.66. To enhance the quality of model outputs, we employ in-context learning (Brown et al., 2020). For each step, due to the limit of context length, we construct 2 human-annotated examples for in-context learning. We ask LLMs to regenerate for fragmentary outputs.

B Details of Evidence Annotation

Following previous work (Huang et al., 2023b), we train a classification model to retrieve relevant articles. We fine-tune RoBERTa-large (Liu et al., 2019) with 80K examples. Each example consists of one question and 1~5 articles. For each consultation question, we keep the top 10 articles with the highest probability scores predicted by the classifier.

To avoid the model failing to retrieve articles that should serve as the basis, we employ 6 annotators with legal backgrounds to supplement missing articles. They are then asked to assess whether each article should/can serve as the basis to respond the given question based on the following principles:

- **Necessary:** The article is highly relevant to the question. Without this article, it is impossible to answer the question.
- **Optional:** The article pertains to potential scenarios that may be relevant to the question. This article can be used for extended discussions about the scenarios that user may concern or face to in near future.
- **Not required:** The article is entirely unnecessary to address the question.

We assign relevant scores of 2, 1, and 0 to the three categories of articles respectively. If the average score of an article exceeds 1.66, it will be regarded as a *necessary* one. And the articles with average scores less than 0.67 can be regarded as *not required*, while the remaining ones are *optional*.

To imitate retrieval-augmented generation, we provide five articles for each question, supposing them as retrieval results. We first keep all the *necessary* and *optional* articles. We then select *not required* articles in descending order of the probability scores predicted by the classifier.

C Rules to Identify the Article Reference

We examine whether the responses use an article as the basis by following rules:

Model	Avg. Lens. of base	Avg. Lens. of CoD
Baichuan2-7B	201.20	200.78
Deepseek-7B	214.29	219.09
Qwen-7B	236.55	259.03
Xverse-7B	236.61	244.60

Table 5: The average lengths of responses generated by different LLMs in baseline setting and CoD setting.

- If the article number appears in the response, we believe the LLM has used this article as a reference.
- We segment the responses into sentence (Che et al., 2021) and calculate the longest common subsequence (LCS) between each sentence and the article's content. If the length of the longest LCS exceeds one-third of the article, we believe the LLM has referenced this article.
- Otherwise, the article is considered not to serve as a reference.

D More Results and Analysis

D.1 Lengths of Generated Responses

In Table 5 we provide the average lengths of generated responses from each LLM in different settings. We find that the lengths of responses under different experimental settings can be similar. After checking the correlation coefficient between the difference in response lengths and the difference in GPT-4 scores, we believe that the GPT-4 evaluator does not show preference for longer responses.

D.2 Examples required Reasoning with Commonsense Knowledge

As shown in the *original analysis* of Table 6, LLMs tend to focus on the literal differences between questions and law articles. We humans have the background knowledge that only courts have the authority to revoke guardianship, thus often omit this information in our questions. But, LLMs often struggle with such differences between the questions and rigorous law articles, thus may not yield correct analysis.

D.3 Significant Test

We apologize that we have not provided significance test. We have to admit that the GPT-4 evaluation is indeed too expensive. A single evaluation of all results under all settings can cost $150. It is difficult for us to repeat the evaluation multiple times, since it may cost thousands of dollars.

In this work, we use a low temperature when generating. And we set the temperature to 0.0 for GPT-4 evaluation.

We randomly sample 20 examples and employ Baichuan2-7B to repeat the experiments of the baseline setting for 5 times with different seeds. We then use GPT-4 to score the responses of every run. The standard error of the average scores is 0.0245, which is much smaller than the improvement brought by CoD.

We also repeat the experiments on Baichuan2-7B for 5 times under the setting of CoD, using the same subset. The standard error of average scores yielded by GPT-4 is 0.0510, which is also smaller than the improvement brought by CoD.

Question: Do you still need to pay child support after having your guardianship revoked?
Article: *Article 37* Parents, children, and spouses who support the wards in the form of child support, support for elderly parents, or spousal support in accordance with the law shall continue to perform such obligations after they are disqualified by the people's courts as guardians.
Original analysis: *Article 37* explicitly stipulates [the content of Article 37]. This article pertains to guardianship and child support, but since the question does not mention revocation by the People's Court, this article should not be used as a basis.

Low-quality modification:
Revised analysis: *Article 37* explicitly stipulates [the content of Article 37]. This article pertains to guardianship and child support. However, the user does not explicitly say who revokes her/his guardianship. Thus, this article should not be used as a basis.

High-quality modification:
Revised analysis: Article 37 stipulates that the revocation of guardianship does not affect existing obligations to pay child support. Thus, this article should be used as a basis.

Table 6: Failed and successful cases for revising evidence analysis. Red texts are the key basis of the question. Violet texts are correct analysis, while the texts with yellow background are hallucinated parts.

E Prompts of Chain-of-Discussion

E.1 Prompt of Question Analysis

To obtain the question analysis, we employ the prompts as below ([]indicates the English translation of the prompts in Chinese):

你是一个民法领域的专家，你需要从法律专业的角度分析一名咨询者提出的问题涉及哪些关键点。在分析问题之后，你还要分析检索器提供的参考法条是否能作为分析该问题的依据。请你紧紧围绕咨询者的问题进行分析，不要过度设想潜在的、与问题不相关的场景。

[*You are an expert in the field of civil law. You need to analyze the key points of a question posed by a consultant* *from a legal professional's perspective. After analyzing the question, you should also assess whether the reference articles provided by the retriever can serve as a basis for analyzing the issue. Please focus closely on the consultant's question and avoid overly imagining potential scenarios unrelated to the issue.*]

咨询者的问题是"[[question]]"

[*The consultant's question is:* "[[question]]"]

下面是检索器提供的参考法条：[[articles]]

[*Below are the reference articles provided by the retriever:* [[articles]]]

接下来，请你分析咨询者的问题"[[question]]"请你站在咨询者的角度，首先分析咨询者所处的环境及其面对的客观事实，再围绕着咨询者的问题，分析他可能需要了解哪些方面的法律规定。你需要先对问题进行分析，再分析各个参考法条是否有该问题有关。请你遵循格式，以"问题分析："、"法条分析："分别作为两段分析的开头。

[*Next, please analyze the consultant's question* "[[question]]" *Please put yourself in the consultant's shoes, first analyzing the environment they are in and the objective facts they are facing. Then, focusing on the consultant's question, analyze which aspects of the legal articles they might need to understand. You need to first analyze the question, then assess whether each reference article is relevant to the issue. Please follow the format, starting each section of the analysis with "Question Analysis:" and "Article Analysis:" respectively.*]

Since we find the LLMs tend to provide the article analysis accompanied with question analysis. We ask it to generate these two analyses separately. We stop the generation after "法条分析：" ([Article Analysis]) has been output.

E.2 Prompts of Summary

To summarize the question analyses from multiple LLMs, we use following prompt:

你是一名法律专家，你需要评价数位律师针对某个法律咨询问题的分析，并给出最终的总结性分析。这些律师的分析既包含正确的内容，也包含错误的内容。你可以参考法条中的内容，谨慎判断各位律师对问题本身的分析是否正确，以及对潜在情况的讨论是否合理。请注意，多数律师都提及的内容更有可能是正确的，你在最终总结时可以参考这部分内容。而如果某个内容仅被个别律师提及，那么该内容有可能是错误的或者与问题不相关的，你需要谨慎判断它是否合理。

[*You are a legal expert. You need to evaluate several lawyers' analyses of a specific legal consultation question and provide a final summary analysis. These lawyers' analyses contain both correct and incorrect elements. You can refer to the content of the legal articles, carefully judging whether each lawyer's analysis of the issue itself is correct and whether their discussion of potential scenarios is reasonable. Note that content mentioned by most lawyers is more likely to be correct, and you can reference this part in your final summary. If a particular point is mentioned by only a few lawyers, it may be incorrect or irrelevant to the issue, so you need to carefully judge its reasonableness.*]

你需要先对律师们的分析进行点评，你需要评价这些分析中的每一个要点是否正确、是否与问题相关。对于存在逻辑错误或者与问题相关度较低的要点，你要明确指出并给予批评。之后，请你基于你的点评，给出一段语气、句式都与各位律师对该问题的分析相似的总结性分析。

[*You first need to critique the lawyers' analyses, evaluating whether each point in these analyses is correct and relevant to the issue. For points with logical errors or low relevance to the issue, you should clearly point them out and provide criticism. Afterward, based on your critiques, provide a summary analysis that matches the tone and style of the lawyers' analyses of the issue.*]

E.3 Prompts of Evidence Analysis

To obtain analyses for articles, we use following prompts:

你是一名法律专家，你需要判断某个法条是否能作为依据，用于解答给定的法律咨询问题。针对咨询者提出的问题，检索器提供了五个参考法条。但是这些法条可能对于解答问题有帮助，也可能没有帮助。你需要逐个对法条进行分析和判断，在针对某个法条进行判断时，请你不要对其他法条进行判断。

[*You are a legal expert tasked with determining whether a specific legal article can serve as a basis for answering a given legal consultation question. The retriever has provided five reference articles for the question posed by the consultant. However, these articles may or may not be helpful in answering the question. You need to analyze and evaluate each article individually, refraining from making judgments about other articles when assessing a particular one.*]

咨询者的问题是"[[question]]"

[*The consultant's question is:*[[question]]]

检索器提供的参考法条：[[articles]]

[*The reference articles provided by the retriever are:* [[articles]]]

你需要先对该问题的关键点进行分析，然后再逐个分析每个法条是否对于解答该问题有帮助。分析法条的过程中，请你先思考法条规定了何种权利和义务，或者对何种行为实施了禁止令。如果法条中规定或禁止的内容与问题中的关键点有一定的相关性，那么该法条有可能对于解答问题有所帮助；否则，该法条大概率对解答问题没有帮助。

[*You need to first analyze the key points of the question and then proceed to analyze each article to determine its relevance to answering the question. During the analysis of the articles, consider what rights and obligations the article specifies, or what behaviors it prohibits. If the articles are somewhat relevant to the key points of the question, then the article may likely be helpful in answering the question; otherwise, the article is probably not helpful in addressing the question.*]

E.4 Prompt of Critique

When criticize other LLM's article analysis, we use following prompts:

你是一名法律专家，你需要点评一名律师对于某个法条是否有助于解答某个法律咨询问题的分析是否误解了法条的内容。我会明确告知你问题和法条的具体内容。

[*You are a legal expert. You need to critique a lawyer's analysis of whether a specific legal article is helpful in answering a particular legal consultation question and determine if their analysis misconstrues the content of the legal article. I will provide you with the specific details of the question and the legal article.*]

问题：[[question]]

[*Question:* [[question]]]

法条：[[article]]
[*Articles:* [[article]]]
律师对于法条的分析：[[art_ana]]
[*Lawyer's analysis for the articles:*[[art_ana]]]
接下来，请先用简洁的语言点评律师对于[[cur_art_id]]的分析。之后，请你判断他的分析是否误解了法条的内容。

[*Next, please provide a concise critique of the lawyer's analysis of* [[cur_art_id]]. *Then, determine whether their analysis misconstrues the content of the legal article.*]

E.5 Prompt of revising

When revising, we employ following prompts:
你是一名律师，你对于某个法条是否有助于解答某个法律咨询问题进行了点评。一些法学专家认为你的点评中存在对法条内容的理解、法条与问题之间的关联性等角度存在错误。你需要参考你对问题的分析，修改你对法条的分析。
[*You are a lawyer who has provided an assessment of whether a certain legal article is helpful in addressing a specific legal consultation question. Some legal experts believe there are errors in your assessment regarding understanding the content of the legal article and its relevance to the question. You need to revise your analysis of the legal article based on your analysis of the issue.*]
问题：[[question]]
[*Question:*[[question]]]
法条：[[article]]
[*Articles*[[article]]]
律师对于问题的分析：[[que_ana]]
[*Lawyer's Analysis of the Question:* [[que_ana]]]
律师对于法条的分析：[[art_ana]]
[*Lawyer's Analysis of the Legal Article:* [[art_ana]]]
专家点评：[[critiques]]
[*Expert Critiques:* [[critiques]]]
接下来，请你重写一份更为正确的法条分析。在重写后的法条分析的结尾，请你按照你的分析，评估一下该法条是否可能有助于解答问题。

[*Next, please rewrite a more accurate analysis of the legal article. At the end of the rewritten analysis of the legal article, evaluate whether the legal article may indeed be helpful in addressing the question based on your analysis.*]

E.6 Prompt of response

Finally, to response to the user's question, we use following prompts:

你是一个法律专家，你需要从法律专业的角度回答咨询者提出的问题。你需要以具体的法条为依据回答问题，并告诉咨询者法律赋予他哪些权利和义务，或者禁止他实施哪些举措。在回答问题之前，你可以参考检索器提供的一些参考法条。但请注意，检索器提供的法条并不一定都有助于回答咨询者提出的问题，它也可能与提问者的问题无关。因此，你需要对问题涉及的事实背景进行分析，再判断各个法条是否能够作为回答问题的依据。请你不要将检索器提供的全部参考法条都当作依据，也不要引用参考法条之外的其他法条作为依据。在回答的过程中，请你紧紧围绕提问者的问题进行讨论，不要过度设想潜在的、与问题不相关的情形。

[*You are a legal expert, and you need to answer the question posed by the consultant from a legal perspective. You are required to provide specific legal articles as the basis for your answer, informing the consultant of their rights, obligations conferred by the law, or actions prohibited by it. Before answering the question, you can refer to some reference articles provided by the retriever. However, please note that the articles provided by the retriever may not necessarily be helpful in answering the consultant's question; they may also be irrelevant to the question. Therefore, you need to analyze the factual background of the issue involved, then determine whether each article can serve as a basis for answering the question. Please do not consider all the reference articles provided by the retriever as the basis, nor cite any articles outside the reference ones as evidence. During your response, focus closely on the consultant's question, avoiding overly imagining potential scenarios unrelated to the issue.*]

咨询者的问题是"[[question]]"

[*The consultant's question is:* "[[question]]"]

下面是检索器提供的参考法条：[[articles]]

[*Below are the reference articles provided by the retriever:* [[articles]]]

接下来，请你回答咨询者提出的问题"[[question]]" 你需要先对该问题的关键点进行分析，然后判断各个参考法条是否有助于解答该问题。最后请你使用与该问题有关的部分法条作为依据，给出详细的回答。回答过程中禁止使用参考法条之外的内容。

[*Next, please answer the question posed by the consultant* "[[question]]" *You need to analyze the key points of the question first, then determine whether each reference article is helpful in answering it. Finally, please provide a detailed answer using relevant portions of the articles as the basis. Use of content outside the reference articles is prohibited during the response.*]

问题分析：[[que_ana]]

[*Question Analysis*: [[que_ana]]]

法条分析：[[art_ana]]

[*Article Analysis*: [[art_ana]]]

回答：

[*Answer:*]

F Scoring Prompt of GPT-4

Following CritiqueLLM (Ke et al., 2023), we employ a reference-based prompt to instruct GPT-4 to assess the overall quality of the responses generated by open-source LLMs. We use the human-written response and the *necessary* and *option* articles as reference. The prompt is shown as below:

[Instruction]

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the logicality, helpfulness, relevance, accuracy, depth, and whether using irrelevant articles beyond the reference articles as a basis. Begin your evaluation by providing a short explanation. You will be given several reference articles, a high-quality reference answer and the assistant's answer. Be as objective as possible. You should first provide your explanation IN CHINESE, then you must rate the response on a scale of 1 to 10 by STRICTLY following the below MAPPING for the relation between the scores and response quality:

1. The score 1~2 stands for very chaotic or absence of answer, and the AI assistant completely failed to answer the user's question. Serious logical and factual errors might also be included in this term. The gap between the AI assistant's answer and the high-quality reference answer is huge and insuperable.

2. The score 3~4 indicates fragment-like responses from AI assistant's answer. It did not provide answers in proper grammar, fluency, or accuracy. Citing irrelevant articles and resulting in a redundant output also falls under this scenario. There are obvious gaps between the high-quality reference answer and the AI assistant's response.

3. The score 5~6 indicates for existence of minute disadvantage from the AI assistant's answer compared to the high-quality reference answer. Yet the AI assistant did provide an average answer. The AI assistant either did not fully address the question, or was somewhat short of logicality, helpfulness, relevance, depth, or detailedness. The disadvantages from the AI assistant's answer overwhelm its advantages.

4. The score 7~8 indicates the AI assistant provided a good answer as well as the high-quality reference answer, addressing the question, with good helpfulness, relevance, accuracy, depth, creativity, and enough details. The response of AI assistant does not include any irrelevant articles beyond the reference articles. The AI assistant might have flaws compared to the reference answer, but that does not overwhelm the above advantages.

5. The score 9~10 indicates the AI assistant responded better than the provided reference answer in most aspects, fully achieved the instruction goal, provided more detailed analysis, and have unique advantages to the reference answer.

You should give scores around 7 if you do not find obvious advantages or disadvantages. You should seriously consider the above guide before give lowest and highest scores such as 1 or 10, and avoid such situation if you do not have sound explanations. Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. AND again, VERY IMPORTANTLY, after you provide your explanation IN CHINESE, you must rate the response strictly following this format: "Rating: [[Number]]", for example: Rating: [[5]].

[User's Question]
{{QUESTION}}
[The Start of Reference Articles]
{{ARTICLES}}
[The End of Reference Articles]
[The Start of Reference Answer]
{{GOLDEN RESPONSE}}
[The End of Reference Answer]
[The Start of Assistant's Answer]
{{LLM'S RESPONSE}}
[The End of Assistant's Answer]