

# 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

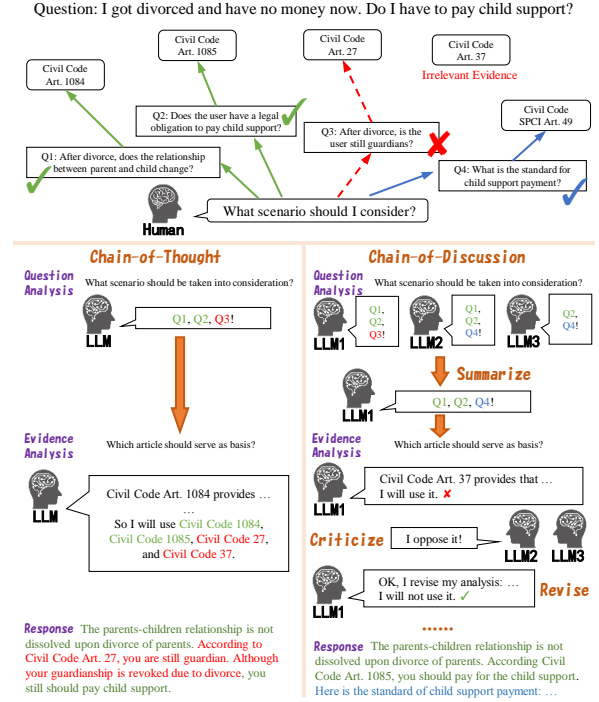


Figure 1: The process of Chain-of-Discussion, compared with chain-of-thought. The green parts are necessary to answer the user’s question. The blue parts indicate closely related to the question, which can be used for detailed discussion. The red parts are irrelevant content which should be avoided.

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 perfect, inevitably with noise in the retrieval results. Let us take legal consultation as an example. 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 the LLMs output correct responses and consistent explanations, while providing more useful suggestions about the potential scenarios not directly mentioned in the questions but indeed helpful for users’ current or near-future situations. For instance, 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 in the burden of child support given her/his current situation. This is even hard for humans where one should have access to proper evidence, e.g., the necessary or closely related law articles here, and accordingly provide kind reminders with reasonable explanations. Let alone LLMs without abundant annotations to train/fine-tune, which usually focus on the specific facts literally appearing in the questions.

In this work, we will 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 above challenges, we propose a novel chain-of-thought framework, the Chain-of-Discussion (CoD), which involves multiple LLMs in summarizing, criticizing and revising each other’s output to reach a well-supported and helpful response.

Our motivations are two-fold. On the one hand, 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 possibil-

ities. 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 APIs, 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 challenges more complex 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 is irrelevant evidence is also a challenging issue for humans.

**Chain-of-Thought (CoT)** 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  $\mathcal{D}_q = \{d_1, \dots, d_t\}$  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, judicial interpretations, 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 the evidence out of context. (2) **Comprehensiveness**: The responses should engage in discussions about potential scenarios that would be relevant or helpful to the users, even if not explicitly mentioned in the users’ 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  $a_q^{\text{que}}$ .

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

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

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, at the stage of question analysis, we encourage models to read and summarize others’ analyses so as to take more scenarios closely relevant to the question into account, 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  $M_1, \dots, M_n$ . 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:  $a_{q, M_i}^{\text{que}} = f_{\text{que}}(q, \mathcal{D}_q | \theta_{M_i})$ .

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  $a_q^{\text{que}} = f_{\text{sum}}(q, a_{q, M_1}^{\text{que}}, \dots, a_{q, M_n}^{\text{que}} | \theta_{M_k})$ .

#### 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:  $\hat{a}_{d_j}^{\text{evi}} = f_{\text{evi}}(d_j, q, a_q^{\text{que}} | \theta_{M_k})$ .

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  $\hat{a}_{d_j}^{\text{evi}}$ . Each critic model  $M_i$  should explicitly output whether it holds opinions contrary to  $\hat{a}_{d_j}^{\text{evi}}$ , which are denoted as  $c_i^{d_j}$ . In this work, we employ a revising threshold  $\delta$ . If the proportion of opposite opinions in the critiques exceeds  $\delta$ , the target model needs to **revise** its evidence analysis:  $a_{d_j}^{\text{rev}} = f_{\text{rev}}(q, d_j, a_q^{\text{que}}, \hat{a}_{d_j}^{\text{evi}}, \{c_i^{d_j}\}_i, \theta_{M_k})$ .

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:

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

#### 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  $r = f_{\text{ans}}(q, \mathcal{D}_q, a_q^{\text{que}}, \{a_{d_1}^{\text{evi}}, \dots, a_{d_t}^{\text{evi}}\} | \theta_{M_k})$ .

### 5 Experiments

As discussed in Section 3, legal consultation is a typical example for CEBQA tasks, which require model to generate an **accurate** response including **helpful** discussions about relevant scenarios. In our experiments, we delve into the legal consultation task in China, where all legal activities should be based on *law articles* and *judicial interpretations*, which can be naturally considered as the evidence store in our framework.

#### 5.1 Data Collection

We focus on the legal consultation in the fields of *marriage, family affairs, and inheritance*, which cover various types of legal disputes such as divorce, custody, contracts, property and so on. We



collect 200 questions from real users and the corresponding responses from consultants through Web Search Engines.

**Data Quality** To ensure the data quality, we manually check on the questions and answers. We correct all typos but retain the informal expressions in the questions. Note that there may be omissions or slight word-order inversions in the questions, which poses a challenge to the model’s reasoning capabilities.

We employ two annotators with background in civil law to examine the correctness and logical coherence of these responses. For the responses identified with errors, we encourage the two annotators to discuss and reach a consensus for modifications, otherwise, leave them as they are.

**Evidence Annotation** We construct the evidence store based on all articles of the *Civil Code* and the *Civil Procedure Law* and their judicial interpretations. We categorize these article into three types: *necessary*, *optional*, and *not required*. The *necessary* articles are the ones highly relevant to the question, while the *optional* articles can be basis for the discussion of potential scenarios. Please see more details in Appendix B. We ensure there are 5 articles in each example. And on average, each example contains 1.52 *necessary* articles, 1.23 *optional* articles, and 2.25 *not required* articles. It means approximately 45% of the retrieved articles are not required at all.

## 5.2 Experimental Setup

In this work, we select open-source LLMs trained by different research groups. We hope these models have learnt different knowledge and gain different reasoning capabilities from pre-training. Then, these models may provide various analytical perspectives via interaction and compensate for deficiencies in their own reasoning capabilities.

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<sup>1</sup>, 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.

<sup>1</sup><https://huggingface.co/xverse/XVERSE-7B-Chat>

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):** Question analysis, article analysis, and response are all generated by a single LLM.

**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 target LLM. The question analysis and the final response are generated by 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. Please see more details in Appendix A.

**Evaluation Metrics** Different from the short-form open-domain QA whose answers are usually several words or sentences, the responses in the CEBQA tasks can consist of several hundred or even thousands of words. Therefore, it is impossible to employ the metrics such as F1 or exact match which are widely used in QA benchmarks (Joshi et al., 2017; Rajpurkar et al., 2018). These generated responses are also complex which contain facts and causal relations to be verified. Thus, it is will be difficult to evaluate them all by humans due to unacceptable time costs.

Following previous works (Liu et al., 2023; Chan et al., 2024), we employ GPT-4 to evaluate the quality of generated responses, with the human-written responses, necessary and optional articles as reference. We prompt gpt4-turbo-0125 to score the responses based on correctness and comprehensiveness. The scorer should assign integer scores between 1 and 10. If there is no clear reason to indicate that the responses of LLMs are significantly better or worse than human-written ones, a score of around 7 should be given. Please see the scoring prompts in Appendix E.

## 5.3 Main Results

Table 1 shows the evaluation results produced by GPT-4. Comparing the results of baseline framework (BS) and Chain-of-Discussion (S1S2), we can

Target LLM	Setting	Avg. Score	$\Delta$ Score
Baichuan2-7B	BS	5.750	–
	S1	6.030	+0.280
	S2	5.935	+0.185
	S1S2	<b>6.090</b>	+0.340
Deepseek-7B	BS	6.465	–
	S1	6.505	+0.040
	S2	6.480	+0.015
	S1S2	<b>6.580</b>	+0.115
Qwen-7B	BS	5.835	–
	S1	5.890	+0.055
	S2	5.815	-0.020
	S1S2	<b>5.955</b>	+0.120
Xverse-7B	BS	6.015	–
	S1	5.995	-0.020
	S2	6.030	+0.015
	S1S2	<b>6.125</b>	+0.110

Table 1: The average scores of each target LLM and each setting evaluated by GPT-4.

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 also find that employing multi-model discussion on both stages can bring more improvement than using it on one stage only.

We also have to acknowledge that although Chain-of-Discussion can enhance the LLMs, the CoD-augmented Baichuan2-7B, Qwen-7B, or Xverse-7B can still not outperform Deepseek-7B under baseline settings, with around 0.5 scores left behind. The results show that the quality of responses primarily relies on the inherent ability of LLM to comprehend contexts and then to generate.

We notice using multi-model discussion only at Stage 1 or Stage 2 fails to enhance Xverse-7B or Qwen-7B. We will provide more discussions and case studies in Section 6.3.

## 6 Discussions

### 6.1 Evidence-Centric Evaluation

Different from previous Question Answering tasks (Joshi et al., 2017; Kwiatkowski et al., 2019) whose answers can be several words or single sentences, in the CEBQA tasks, LLMs are required to provide both detailed and correct responses for the question and potential scenarios. Therefore, we wonder whether the Chain-of-Discussion framework can enhance the comprehensiveness and correctness of model output.

Similarly, when discussing the details of potential scenarios, LLMs should also reference *optional*

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

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.

evidence. Hence, we can assess the correctness and comprehensiveness of responses by the accuracy of reference to various types of evidence documents.

We propose 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 rule-based method to examine whether the response have used an article. Please see details in Appendix C.

We compute the Macro average N-Acc and O-Acc across all examples. If an example does not contain *optional* articles, it will not participate in the calculation of O-Acc. Table 2 shows the results of each target LLM under different experimental settings. Compared to the baselines (BS), the Chain-of-Discussion framework (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, which performs the best in GPT-4-based evaluation, our proposed framework still brings improvements of 0.86% and 0.53% to N-Acc and O-Acc, respectively. Recalling the GPT-4-evaluated results in Table 1, where Baichuan2-7B obtains the most improvement of overall quality. We find this LLM also get the most improvement on N-Acc and O-Acc, with improvements of 4.91% and 2.24%, respectively.

The results indicate that introducing multi-model discussions during both question analysis and evidence analysis **contributes to increasing the prob-**

**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 parties 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 parties need to do so voluntarily.

Table 3: A case of Qwen-7B obtaining improvement from CoD. Violet texts are correct analysis, while texts with yellow background are hallucinated parts.

**ability of LLMs referencing correct evidence.** It can be one of the reasons why Chain-of-Discussion can improve the quality of model responses.

Comparing the results under the setting of BS, S1, and S2, we can find that involving multiple LLMs in a single stage can actually enhance both correctness and comprehensiveness. However, overall, employing multi-model discussions in question analysis contributes more to the comprehensiveness, while introducing other models in evidence analysis brings more improvement in correctness.

## 6.2 Manual Check

To further study the quality of responses generated by CoD, we randomly sample 30 cases and manually examine the responses in terms of fluency and logic. We select the responses generated by Qwen-7B or Qwen7B+CoD, which get worse average GPT-4-evaluated scores than responses of other LLMs. We find Qwen-7B with vanilla CoT has a poor ability to comprehend the articles, often resulting in logical errors in responses, while the CoD mechanism introduces opinions and critics from other LLMs, thus helps to distinguish ambiguous terms, and reach better logic.

Table 3 show the case where CoD can help to reduce logical errors in Qwen-7B’s responses. We find Qwen-7B fails to understand that both 复婚 (reconcile and remarry) and 再婚 (remarry with other person) are considered as marriage in legal terms, which should comply with the provisions of Article 1046 but not Article 1083. We believe that discussions with other models can, to some extent, reduce the hallucination caused by the target model’s poor reasoning capabilities.

**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 4: A bad case of hallucinated summary of question analysis. Violet texts are correct analysis, while the texts with yellow background are hallucinated parts.

## 6.3 Limitations of Open-Source LLMs

Recalling the results in Table 1 where using multi-model interaction only at one stage fails to enhance Xverse-7B or Qwen-7B. We have to acknowledge that due to the limitations in parameter size, these models might produce unreliable output or demonstrate inferior capabilities to follow instructions.

We guess there can be two reasons. (1) The target LLM may fail to discern errors in the question analysis of other LLMs, and integrate them into the summary. (2) The target LLM does not always follow instructions, and refuse to modify the errors in evidence analysis.

### 6.3.1 Hallucination Propagation in Summary

At the stage of question analysis, we explicitly instruct the target LLM to assess whether the question analyses of other LLMs are logically correct and relevant to the question. However, the target model might still incorporate hallucinated analyses into the summary.

Table 4 illustrates a bad case where Baichuan-7B incorporates inaccurate analysis from DeepSeek-7B into the summary. We find that LLMs can correctly identify the primary intention behind user’s

question, i.e., determining whether providing support for parents is a prerequisite for inheriting. However, Deepseek-7B also engages in analysis regarding scenarios where parents should pay support for children. And Baichuan2-7B fails to identify that this part of analysis is irrelevant to the question.

For people with legal background, it is a commonsense that raising children and caring for elderly parents are two different concepts. And legally, there is no causal relationship between them. However, we guess that for LLMs, both raising children and supporting parents represent financial support and life care between two generations. We believe that open-source LLMs with around 7B parameters may not have sufficient knowledge and enough reasoning capabilities. They may judge whether a question analysis is relevant to the question based on shallow semantic similarities.

### 6.3.2 Model Preference of Revising

Similar to the discussion on question analysis, the lack of good reasoning capabilities may also limit the performance of LLMs in revising evidence analysis. We find that LLMs sometimes tend to focus on the literal differences between questions and law articles, illustrated as the *original analysis* in Table 5. We humans have the background knowledge that only courts have the authority to revoke guardianship, while the LLMs cannot yield correct analysis. Users often omit commonsense information when posing questions, whereas law articles exhibit a rigorous and detailed structure. The stylistic differences between questions and evidence may lead to the misjudgments of LLMs.

We also find that different LLMs have varying preferences regarding whether to modify the evidence analysis. For instance, compared to Qwen-7B, Baichuan2-7B exhibits a greater preference to revise the analyses regarding articles.

To quantitatively study this issue, we initially employ Baichuan2-7B to generate the original evidence analyses. Deepseek-7B and Xverse-7B then play the role of critics to determine which analysis should be revised. Finally, 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 on the conclusions, we also use Qwen-7B to provide original evidence analyses, still with

**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 5: 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.

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 result in the Chain-of-Discussion framework not bringing enough 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 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.

## 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, Sheng-guang 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*.



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<sup>3</sup>: 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  $\delta$  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:

- 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 (Chen et al., 2021) and calculate the longest common subsequence (LCS) between each sentence and the article’s content. If the length

953 of the longest LCS exceeds one-third of the  
954 article, we believe the LLM has referenced  
955 this article.

- 956 • Otherwise, the article is considered not to  
957 serve as a reference.

## 958 D Prompts of Chain-of-Discussion

### 959 D.1 Prompt of Question Analysis

960 To obtain the question analysis, we employ the  
961 prompts as below:

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

967 咨询者的问题是“[[question]]”

968 下面是检索器提供的参考法条：[[articles]]

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

976 Since we find the LLMs tend to provide the ar-  
977 ticle analysis accompanied with question analysis.  
978 We ask it to generate these two analyses separately.  
979 We stop the generation after "法条分析：" has been  
980 outputed.

### 981 D.2 Prompts of Summary

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

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

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

### 999 D.3 Prompts of Evidence Analysis

1000 To obtain analyses for articles, we use following  
1001 prompts:

1002 你是一名法律专家，你需要判断某个法条是否能作为  
1003 依据，用于解答给定的法律咨询问题。针对咨询者提出  
1004 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。

咨询者的问题是“[[question]]”

检索器提供的参考法条：

[[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。

### D.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。

问题：[[question]]

法条：[[article]]

律师对于法条的分析：[[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。

### D.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。

问题：[[question]]

法条：[[article]]

律师对于问题的分析：[[que\_ana]]

律师对于法条的分析：[[art\_ana]]

专家点评：[[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。

### D.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。

咨询者的问题是“[[question]]”

下面是检索器提供的参考法条：[[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。

问题分析: [[que\_ana]]

法条分析: [[art\_ana]]

回答:

## E 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 logicity, 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 logicity, helpfulness, relevance, depth, or detailedness. The disadvan-

tages 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]