LawLuo: A Multi-Agent Collaborative Framework for Multi-Round Chinese Legal Consultation

Anonymous ACL submission

Abstract

Legal Large Language Models (LLMs) have shown promise in providing legal consultations to non-experts. However, most existing Chinese legal consultation models are based on single-agent systems, which differ from real-world legal consultations, where multiple professionals collaborate to offer more tailored responses. **To better simulate real consultations, we propose LawLuo, a multi-agent framework for multi-turn Chinese legal consultations**. LawLuo includes four agents: the receptionist agent, which assesses user intent and selects a lawyer agent; the lawyer agent, which interacts with the user; the secretary agent, which organizes conversation records and generates consultation reports; and the boss agent, which evaluates the performance of the lawyer and secretary agents to ensure optimal results. These agents' interactions mimic the operations of real law firms. To train them to follow different legal instructions, we developed distinct fine-tuning datasets. We also introduce a case graph-based RAG to help the lawyer agent address vague user inputs. Experimental results show that LawLuo outperforms baselines in generating more personalized and professional responses, handling ambiguous queries, and following legal instructions in multi-turn conversations. Our full code and constructed datasets will be open-sourced upon paper acceptance.

1 Introduction

Since the release of ChatGPT, the development of Chinese Large Language Models (LLMs) has accelerated, resulting in influential models like ChatGLM (Du et al., 2022), LLaMa (Touvron et al., 2023), and BaiChuan (Yang et al., 2023). These models excel in fluent Chinese dialogue and understanding complex user intentions. Additionally, domain-specific LLMs, such as Medical (Yang et al., 2024a; Zhang et al., 2023a), Legal (Zhou et al., 2024; Huang et al., 2023), and Financial LLMs (Zhang and Yang, 2023), have emerged, demonstrating strong capabilities in their respective fields.

Recently, notable Chinese legal LLMs, such as LawGPT (Zhou et al., 2024), Hanfei (He et al., 2023), FuziMingcha (Wu et al., 2023) and Lawyer-Llama (Huang et al., 2023), have emerged. These models leverage large Chinese legal dialogue datasets to fine-tune Chinese base models, endowing them with extensive legal knowledge and the ability to engage in legal consultation dialogues. **However**, they fall short of replicating the collaborative workflows of real law firms, limiting their ability to provide personalized, professional responses, as shown in Figure 1.

To address this, we propose **LawLuo**, a multi-agent framework designed to simulate the operations of a law firm and offer professional legal advice, as shown in Figure 1. This framework consists of four distinct agents: a receptionist, a lawyer selected from the lawyer pool, a secretary, and a boss. The receptionist agent is responsible for assessing a user's intent and assigning a lawyer specializing in the relevant field. The lawyer agent analyzes the user's case and provides responses for each round of the conversation. The secretary agent organizes the entire consultation record and generates a final, personalized, and professional response for the user. The boss agent monitors the performance of both the lawyer and the secretary agents. We design a interaction strategy between these agents to simulate the operational processes of real law firms, enabling seamless collaboration to address users' legal consultations.

To enhance the ability of each agent to follow legal instructions, we have constructed three fine-tuning datasets, including: a dataset comprising (Inquire, Lawyer description) pairs for fine-tuning the receptionist agent, a **MU**lti **R**ounds **LE**gal **D**ialogue (MURLED) dataset for fine-tuning the lawyer agent, and a **L**egal **C**onsultation **R**eport

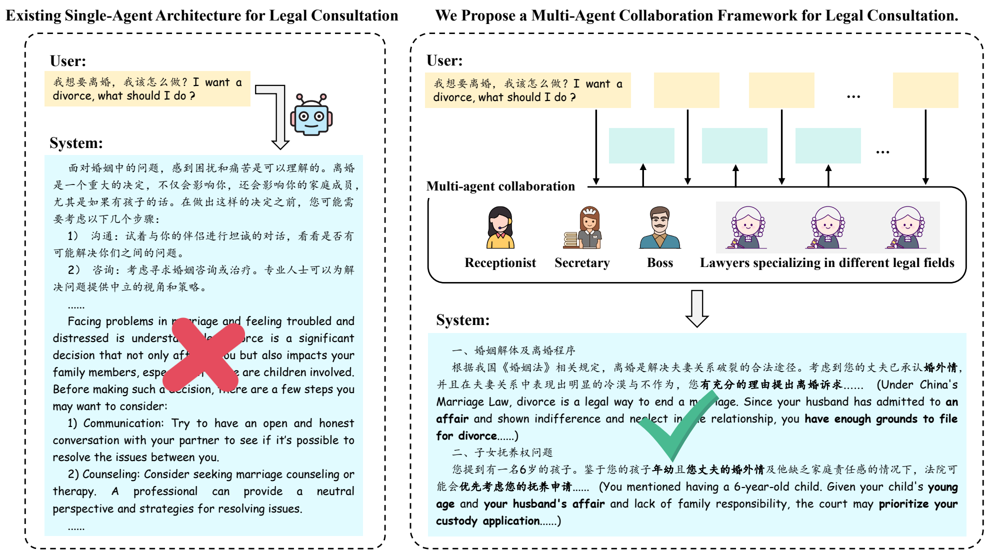


Figure 1: The left side shows the single-agent architecture used by most legal consultation systems, producing superficial, generalized responses without understanding user intent and case details. The right side presents our proposed multi-agent framework, offering more personalized and professional answers.

Generation (LCRG) dataset for fine-tuning the secretary agent. Additionally, to address ambiguous queries, we introduce case graph-based RAG to enhance LawLuo's handling of such queries.

We evaluated LawLuo using GPT-4o and human experts. Experimental results show that LawLuo offers more personalized and professional legal advice compared to baselines. Moreover, when responding to vague questions from users without legal background, baselines often give broad answers directly. In contrast, LawLuo is committed to guiding users to clearly describe case details through leading responses. The experiments also prove LawLuo's strong ability to follow instructions even after multiple rounds of conversation.

Our primary contributions are as follows:

- We introduce a multi-agent collaborative legal dialogue framework that transcends the traditional single-model-user interaction paradigm. This innovation provides users with more personalized and professional consultation services.
- We constructed three different fine-tuning datasets and used them to fine-tune three different agents.
- We propose a case graph-based RAG to handle ambiguous queries from users without a legal background.

2 Related work

2.1 LLMs for Legal Consultation

In recent years, large language models (LLMs) have made significant progress in various fields, particularly in the domain of Chinese law, where they have demonstrated immense potential (Xiao et al., 2021; Yue et al., 2023; Yang et al., 2024b). By training on large volumes of Chinese legal case data, Legal LLMs are able to deeply understand case information and provide users with reasonable legal advice.

Most research relied on continuing pre-training and instruction fine-tuning of existing Chinese base models, aiming to enhance the models' understanding of legal knowledge and their ability to follow legal instructions (Zhou et al., 2024; Huang et al., 2023; Li et al., 2024; Dahl et al., 2024). Their training data mainly consists of publicly available legal documents, judicial exam data, and legal Q&A datasets. Additionally, some studies, such as HanFei (He et al., 2023), have opted to train a legal LLM from scratch, aiming to endow the model with more robust and profound legal knowledge and application capabilities. Some work also utilizes external legal knowledge during the reasoning phase to enhance the model's responses.(Louis et al., 2024; Han et al., 2024; Wan et al., 2024)

However, existing efforts have focused on improving the performance of individual legal LLMs.

In practice, legal consultations in real law firms are often conducted collaboratively by multiple professionals. Inspired by this real-world work model, we propose a multi-agent collaboration framework to simulate this process, thereby providing users with a more personalized and professional legal consultation experience.

2.2 Multi-Agent Collaboration

In LLM-based multi-agent systems, an agent is defined as an autonomous entity capable of perceiving, thinking, learning, making decisions, and interacting with other agents (Xi et al., 2023; Xu et al., 2024). Research shows that breaking complex tasks into simpler subtasks and tackling these with agents that have diverse functions can significantly enhance the problem-solving capabilities of LLMs. (Wang et al., 2024; Guo et al., 2024a). For instance, (Qian et al., 2023) designed a multi-agent collaborative workflow in which agents assuming roles such as CTO, programmer, designer, and tester work closely together to complete software development and document the development process. (Hemmer et al., 2022) have facilitated the construction of machine learning models through collaboration between multiple agents and humans.

In addition, LLM-based multi-agent systems can also be used for simulating real-world social environments, supporting the observation and research of social behavior (Wang et al., 2023; Wei et al., 2023; Du et al., 2023).

We believe that legal consulting is a complex task that should be decomposed into subtasks, which can be collaboratively handled by multiple agents to enhance the personalization and professionalism of the responses.

3 Framework

In real-world scenarios, legal consultations involve collaboration among multiple staff members in a law firm, while current legal LLMs engage with users in isolation. To address this gap, we propose a multi-agent collaborative framework for legal consultation, called LawLuo.

The framework consists of four agent types, as shown in Figure 2: **1)** a receptionist agent, which assesses the user's consultation intent and assigns the appropriate lawyer; **2)** a lawyer agent, selected from the lawyer pool, who interacts with the user to analyze the case details; **3)** a secretary agent, which organizes the dialogue records between the lawyer and the user to generate a final consultation report; and **4)** a boss agent, which monitors the performance of both the secretary and the lawyer to ensure optimal operation.

Given the initial inquiry u_0 from the user, we will now provide a detailed description of the collaborative process of the agents within this framework.

3.1 Receptionist

Given the user's initial inquiry u_0, the receptionist agent R is tasked with evaluating the user's intent and selecting the most suitable lawyer from a pool of candidate lawyers $\mathcal{L}$, each specializing in distinct fields, to address the user's consultation. This process is formalized in Equation 1.

$$R : u_0 \overset{\max}{\mapsto} \arg_{L \in \mathcal{L}} \; similarity(u_0, L) \qquad (1)$$

Where $similarity(\cdot, \cdot)$ represents the similarity between u_0 and the description of lawyer L. We defined 16 descriptions for lawyers specializing in different areas of law, based on the thematic categories of legal consultations on the HuaLv website [1]. These areas include: Contract Law, Labor Law, Corporate Law, Intellectual Property Law, Criminal Law, Civil Procedure Law, Family Law, Real Estate Law, Tax Law, Environmental Law, Consumer Protection Law, Antitrust Law, International Trade Law, Insurance Law, Maritime Law, and others.

We constructed a dataset consisting of 1,600 pairs of (Inquire, Lawyer Description) to fine-tune the Chinese base model BaiChuan (Yang et al., 2023). The fine-tuned model is employed as the receptionist agent R.

3.2 Lawyer

The lawyer agent L, selected from the lawyer pool $\mathcal{L}$, is tasked with engaging in dialogue with the user to acquire a comprehensive understanding of the case details and generate responses. This process is formally represented by Equation 2.

$$L : (u_0, U_{1:T}) \mapsto R_{0:T} \qquad (2)$$

Where $U_{1:T}$ represents the sequence of user queries from the first round to the T-th round, $R_{0:T}$ represents the sequence of the model responses

The existing Legal LLMs, although capable of engaging in dialogue with users, tend to provide one-time responses to user queries. This contrasts

[1] https://www.66law.cn/

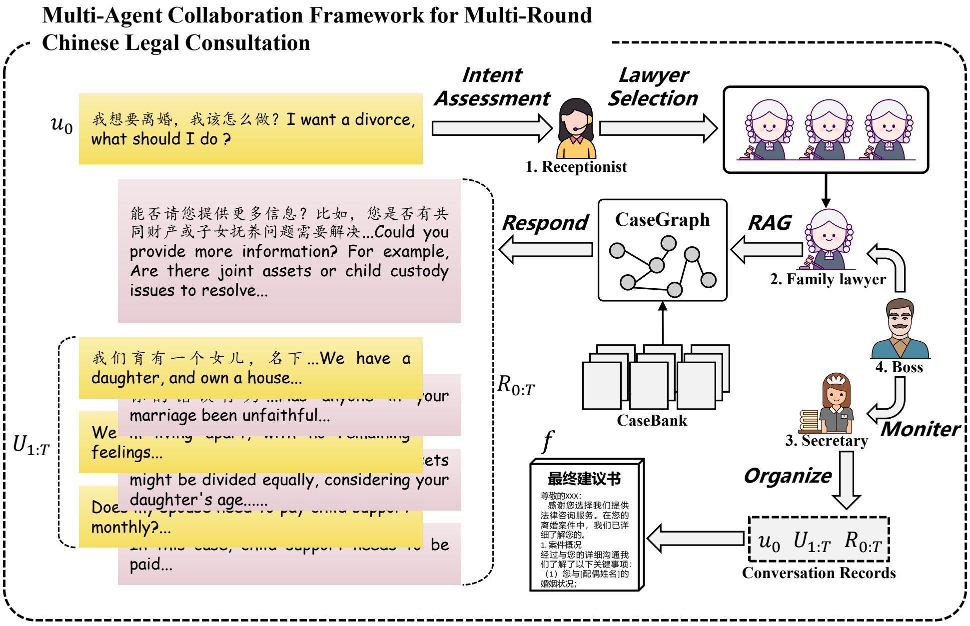


Figure 2: The multi-agent collaboration framework we propose for multi-round Chinese legal consultation. In this framework, the receptionist agent first assesses the user's consultation intent based on the initial input u_0 and selects the most suitable lawyer from the lawyer pool. Subsequently, the selected lawyer agent is responsible for engaging in multi-round dialogues with the user. During this process, the lawyer agent actively queries the user for case details via case graph-based RAG. Finally, the secretary agent organizes the dialogue records between the user and the lawyer, producing a comprehensive consultation report. The boss agent monitors the performance of the lawyer and secretary agents to ensure optimal outcomes.

with real-world legal consultations, where lawyers often engage in multiple guided conversations to gain a deeper understanding of the client's case details. To address this, we constructed a **MU**lti **R**ounds **LE**gal **D**ialogue (MURLED) dataset to fine-tune the Chinese base model ChatGLM (Du et al., 2022), aiming to enhance the model's legal dialogue capabilities, particularly its ability to actively guide in multiple rounds of dialogue. It is worth noting that the MURLED dataset is divided into 16 distinct consulting domains, with 16 different weight checkpoints fine-tuned on Baichuan, each serving as a lawyer agent specialized in a different consulting domain. The distribution of the MURLED dataset across 16 legal consultation fields is shown in Figure 3.

The MURLED dataset was constructed based on case consultation voice recordings from a law firm and contains 16,734 multi-turn legal conversations. We first converted the raw audio files into text format and then utilized ChatGPT to optimize the text, addressing issues such as informality and minor errors. Figure 4 shows an example of a multi-turn conversation from MURLED, highlighting the emphasis on guiding responses.

We used MURLED to fine-tune ChatGLM-3-

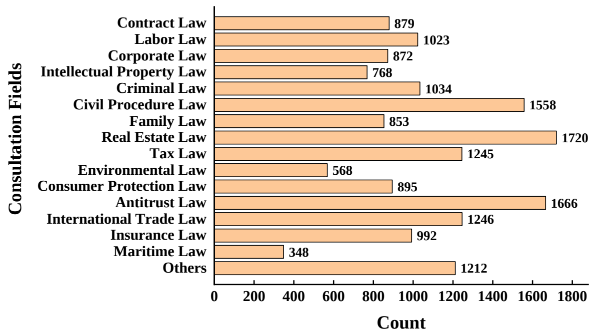


Figure 3: Distribution of the MURLED dataset across 16 different consultation domains.

6b. To mitigate the risk of overfitting, we incorporated general conversational data from Alpaca-GPT4[2], which comprises 52,000 generic Chinese dialogues, into the fine-tuning process. To expedite the fine-tuning of the model and reduce reliance on computational resources, we employed the LoRA fine-tuning strategy, as illustrated in Equation 3:

$$\theta_{Legal} = LoRA\left(\theta, \left\{(u_0^n, U_{1:T}^n, R_{0:T}^n)_{n=1}^N\right\}\right) \tag{3}$$

where θ represents the initial parameters of

[2]https://www.modelscope.cn/datasets/AI-ModelScope/alpaca-gpt4-data-zh/summary

Chinese Sample

```
{
    {
     "Instruction" : "律师您好，我工作期间受伤了，能否申请工伤？",
     "Output" : "为了更好的帮助到您，请告诉我您受伤时间与地点，以及您事故发生的具体原因。"
    }

    {
     "Instruction" : "我是在办公室摔倒的，摔伤了右脚踝，受伤时正好在搬运文件。我已经就医，医生说是轻微扭伤，目前正在休养，预计一周后可以恢复。",
     "Output" : "您的具体伤情如何？是否已经就医？是否有医院的伤情鉴定？"
    }
}
```

English Translation

```
{
    {
     "Instruction" : " I was injured while working, can I apply for work injury compensation?",
     "Output" : "In order to better help you, please tell me the time and place of your injury, as well as the specific cause of your accident."
    }

    {
     "Instruction" : "I fell in the office and injured my right ankle while I was moving documents. I have already seen a doctor, who said it was a minor sprain. I am currently resting and will recover in a week.",
     "Output" : "What are your specific injuries? Have you received medical treatment? Has the hospital issued an assessment of your injuries?"
    }
}
```

Figure 4: An example from the MURLED dataset. It can be seen that this dataset emphasizes the active guidance ability of training large legal models in multi-turn dialogues.

ChatGLM-3-6b, while θ_{Legal} denotes the parameters of our fine-tuned legal LLM. Besides, $(u_0^n, U_{1:T}^n, R_{0:T}^n)$ indicates the n-th training sample.

To enhance the lawyer agent's ability to address vague queries from users without a legal background, we design the agent to employ a case graph-based Retrieval-Augmented Generation (RAG) approach during each response generation process. Specifically, we implement this case graph-based RAG using the LightRAG framework (Guo et al., 2024b). To build the case graph, we utilize a case collection comprising 4,320 criminal cases and 12,345 civil cases, sourced from the China Judgments Online database [3]. The construction of the case graph is detailed in Algorithm 1.

Algorithm 1 Case Graph Construction

Input: Set of legal cases $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$
Output: Case graph $G = (V, E)$
for each case c_i in $\mathcal{C}$ **do**
    $v_i \leftarrow f(c_i)$ ▷ Generate vector representation of case c_i
end for
For each pair of cases:
for each $c_i, c_j \in \mathcal{C}$ **do**
    $Sim(c_i, c_j) \leftarrow \text{similarity}(v_i, v_j)$ ▷ Compute similarity between cases
    Add edge $(c_i, c_j, Sim(c_i, c_j))$ to E ▷ Add weighted edge to graph
end for
Return: Case graph $G = (V, E)$

[3]https://wenshu.court.gov.cn/

3.3 Secretary

The secretary agent's responsibility is to organize the conversation records between the user and the lawyer, and compile a final consultation report to be submitted to the user, as shown in Equation 4.

$$S : (u_0, U_{1:T}, R_{0:T}) \mapsto f \tag{4}$$

Where f represents the final consulting report.

We created a **L**egal **C**onsultation **R**eport **G**eneration dataset called **LCRG**, which includes 420 legal consultation dialogues and their summary reports. Each summary report is carefully written by professional lawyers. We used LCRG to fine-tune the Chinese base model BaiChuan, enabling the model to generate consultation reports from legal consultation dialogues. A legal consultation summary report sample is shown in Appendix A.

3.4 Boss

The boss agent is responsible for evaluating and optimizing the performance of the lawyer and secretary agents. We treat the boss agent as a binary reward model, $B : o \mapsto y$, where o represents the output of the lawyer or secretary agent, and y represents the evaluation of o by the boss agent, categorized as "better" or "worse." The training objective for the boss agent is to minimize the following loss function:

$$\begin{aligned}\mathcal{L}_B = -\frac{1}{N}\sum_{i=1}^{N} [y_i \cdot \log\left(\hat{y}_i(o_i;\theta_B)\right) \\ +(1-y_i)\cdot\log\left(1-\hat{y}_i(o_i;\theta_B)\right)]\end{aligned} \tag{5}$$

where y_i represents the true label of the i-th sample, taking values of either 0 or 1, which correspond to "worse" and "better", respectively. Besides, $\hat{y}_i(o_i;\theta_B)$ denotes the probability that boss predicts the i-th output o_i as "better".

We adopt the PPO algorithm (Wang et al., 2020) to enable reinforcement learning between the boss agent and the lawyer agent, as well as between the boss agent and the secretary agent. Through this reinforcement learning, the boss agent continuously optimizes the lawyer and secretary agents.

4 Experimental Setup

All our experiments were conducted on a 40G A100 GPU. The PyTorch 2.3.0 and the Hugging-Face Transformers 4.40.0 were used. The learning rate for LoRA fine-tuning was set to 0.00005, with a training batch size of 2, over a total of 3 epochs, and model weights were saved every 1,000 steps. Additionally, the rank of LoRA was set to 16, the alpha parameter was set to 32, and the dropout rate was set to 0.05.

5 Results and Analysis

The following research questions will be addressed through experimental analysis:

RQ 1: Does multi-agent collaboration facilitate the generation of more personalized and professional responses for users' legal consultations?

RQ 2: Can LawLuo more effectively address legal inquiries raised by users without a legal background?

RQ 3: After engaging in multiple rounds of legal dialogue, can LawLuo maintain its ability to follow legal instructions accurately?

RQ 4: Does the instruction fine-tuning applied to the constructed datasets improve the performance of the agents?

RQ 5: Is LawLuo still effective in performing routine legal tasks, including non-dialogue tasks?

5.1 Pairwise Comparison Evaluation

We employed pairwise comparison to assess the performance of LawLuo. In the evaluation, the outputs generated by LawLuo is compared with the outputs generated by the baselines using the same input data, in terms of personalization and professionalism, by GPT-4 or human experts. For each comparison, experts are asked to determine whether LawLuo performs better, worse, or similarly to the baseline models. This evaluation method is consistent with current best practices for evaluating large language models (Thirunavukarasu et al., 2023; Xiong et al., 2023; Zhang et al., 2023b).

Figure 5 presents the win rate of LawLuo against the baselines, clearly showing that LawLuo outperforms widely used Chinese base LLMs and exceeds all legal LLM baselines. This results answer **RQ 1**: Collaboration among multiple agents in legal consultation can indeed provide users with more personalized and professional responses.

5.2 Case Study on Ambiguous Inquiry

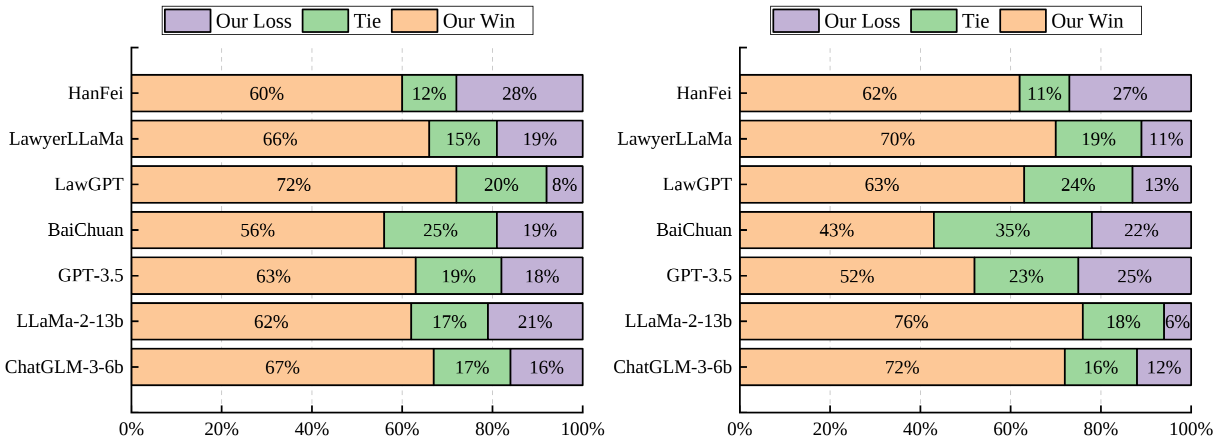
We randomly select a ambiguous legal consultation question and analyze the answers generated by LawLuo, ChatGLM-3, BaiChuan, LawGPT, and HanFei in the first round, as shown in Table 2. From the table, it can be seen that LawLuo's responses in the first round are more guiding. This guiding response helps users to better elaborate on the case details, thereby providing the most personalized and accurate answers. The experimental results address **RQ 2**: LawLuo is better at handling ambiguous legal consultations from users without a legal background.

5.3 Multi-Turn Dialogue

We systematically evaluate the instruction-following capability of the proposed LawLuo model in multi-turn dialogues. The experimental design includes four dialogue scenarios where instructions evolve or become progressively more complex across turns. We assess the model's ability to understand and execute instructions through tasks such as legal charge prediction, similar case matching, and case element extraction. The evaluation metrics focus on the model's accuracy in understanding instructions, coherence in maintaining context, precision in execution, adaptability, flexibility, and its ability to handle complex or conflicting instructions. We use GPT-4 to evaluate LawLuo and the baselines' instruction-following scores after multiple rounds of dialogue, as detailed in Appendix C. The experimental results are shown in Figure 6, with the pink line representing LawLuo's instruction compliance score. It can be observed that even after five rounds of dialogue, LawLuo still maintains a high level of instruction compliance. This experimental result answers **RQ3**: LawLuo is still able to effectively comply with legal instructions after multiple rounds of dialogue.

5.4 Ablation Study

This section aims to validate the contributions of each component within the framework. We continue to use GPT-4o as the evaluator to assess the win rate of LawLuo over GPT-3.5 after ablation, as illustrated in Figure 7. From the figure, it is evident that the win rate of LawLuo over GPT-3.5 decreases by 2% after ablating the receptionist agent. This result validates our hypothesis that legal LLMs should be assigned different domain-specific roles to provide more targeted answers based on the user's consultation field. Additionally, the figure shows that the boss agent also contributes to LawLuo's performance, as it can optimize the responses generated by the lawyer. Finally, we observe a significant decline in model performance



(a) GPT-4 Evaluation (b) Human Evaluation

Figure 5: Win rate of LawLuo compared to the baselines

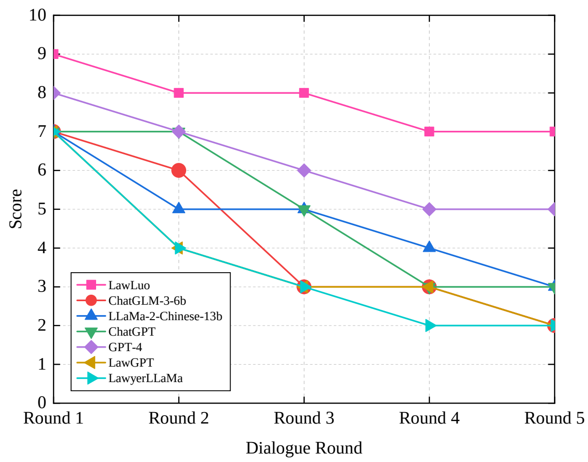


Figure 6: The variation in the quality of model-generated responses with increasing dialogue turns

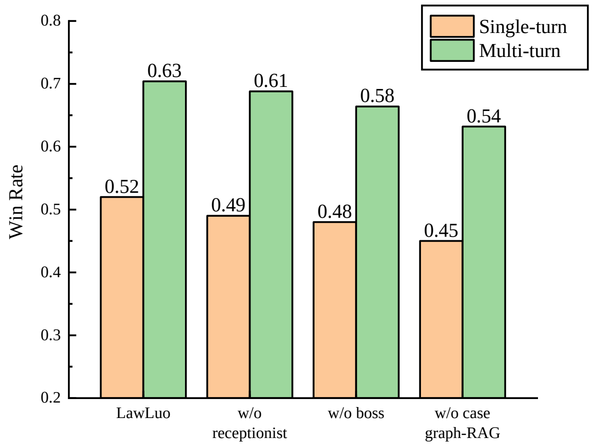


Figure 7: Results of ablation experiments

after removing the case graph-based RAG module. This indicates that clarifying users' vague and ambiguous queries is crucial for generating high-quality responses in legal question-answering. The experimental results answer **RQ4**: Our fine-tuning of each agent enables LawLuo to achieve better overall performance.

5.5 Legal Knowledge Probing Experiment

We evaluated LawLuo's performance across five routine legal natural language processing tasks: Legal Event Extraction, Judicial Reading Comprehension, Legal Charges Prediction, Related Law Retrieval, and Similar Case Retrieval. These tasks were conducted on established datasets, including LEVEN (Yao et al., 2022), CJRC (Duan et al., 2019), CAIL2018 (Xiao et al., 2018), and LeCaRD (Ma et al., 2021).

The results, as presented in Table 1, indicate that LawLuo performs well across all five tasks. Although its performance is not the best when compared to other baseline models, it remains highly competitive. This suggests that through instruction fine-tuning, LawLuo has acquired sufficient legal knowledge, enabling it to not only handle legal consultations but also address routine legal natural language processing tasks. The experimental results answer **RQ5**: LawLuo remains effective in routine legal tasks.

6 System Implementation

Based on the LawLuo framework, we have designed and implemented a practical legal consultation system aimed at providing users with an efficient and interactive legal advisory platform, as shown in Figure 8. The system's backend is developed using the Flask framework, while the frontend is built with React to ensure a dynamic and responsive user experience. Users access the

Table 1: Performance of LawLuo and the baselines on five routine legal natural language processing tasks, reflecting their understanding of legal knowledge.

Task	Legal Event Extraction (F1 score)	Judicial Reading Comprehension (F1 score)	Legal Charges Prediction (F1 score)	Related Law Retrieval (F1 score)	Similar Case Retrieval (Acc@10)
Dataset	LEVEN (Yao et al., 2022)	CJRC (Duan et al., 2019)	CAIL2018 (Xiao et al., 2018)	CAIL2018 (Xiao et al., 2018)	LeCaRD (Ma et al., 2021)
LawLuo	73.3 ± 1.3	80.6 ± 2.3	92.1 ± 2.3	84.4 ± 3.1	81.6 ± 3.8
HanFei	**73.5 ± 1.2**	**83.2 ± 1.7**	92.1 ± 2.2	**84.5 ± 2.0**	83.0 ± 3.1
LawGPT	72.5 ± 1.3	81.5 ± 2.1	90.8 ± 1.5	83.2 ± 2.8	**85.5 ± 3.0**
LawyerLLaMa	71.2 ± 1.4	80.2 ± 2.2	91.7 ± 2.3	81.6 ± 3.1	82.1 ± 3.5
BaiChuan	72.0 ± 1.3	82.2 ± 1.5	92.5 ± 2.3	83.6 ± 3.4	**85.5 ± 2.9**
ChatGLM-3-6b	72.2 ± 1.4	79.8 ± 2.2	**94.4 ± 2.3**	82.1 ± 2.8	84.4 ± 3.3
GPT-3.5	70.9 ± 2.1	77.6 ± 2.4	90.5 ± 2.4	81.1 ± 2.2	80.4 ± 3.2

Figure 8: We have built a web-based legal consultation system with the LawLuo framework as the core, and testing has shown that it has good practical effectiveness.

system through a simple web interface and initially interact with a receptionist agent to describe their legal issues. The system then guides users to the relevant lawyer agent for a detailed case discussion. After several rounds of conversation, the secretary agent generates and provides a legal consultation report, while the boss agent monitors the entire interaction process in the background to ensure service quality.

7 Conclusion

We introduce LawLuo, a multi-agent collaboration framework that simulates the multi-party interactions of real law firms to provide professional legal consulting services. Experimental results demonstrate that LawLuo outperforms traditional single-agent models in generating personalized and professional legal advice, handling ambiguous inquiries, and following legal instructions in multi-turn dialogues. Ablation studies and legal knowledge probing experiments further validate the effectiveness of various components within the framework, as well as the legal knowledge acquired through instruction tuning. Despite these achievements, we acknowledge that there is room for improvement in optimizing inter-agent communication and enhancing model interpretability, which will be the focus of future research. The successful implementation of LawLuo paves the way for new developments in the field of legal consulting, suggesting the broad application prospects of multi-agent collaboration in future legal services.

Limitation and Future Work

The experimental outcomes of the LawLuo framework underscore the potential of multi-agent collaboration within the domain of legal consultation. By emulating the multi-party interactions characteristic of real law firms, our model is capable of delivering consultation service that are more personalized and professional. The strength of this collaborative approach lies in its ability to comprehensively understand user needs from various perspectives and provide solutions on multiple levels. However, multi-agent systems also introduce new challenges, particularly in terms of communication and coordination among agents. To ensure seamless collaboration, each agent must possess a high degree of domain-specific expertise and be able to comprehend the decisions and feedback of other agents. Future work should further explore how to optimize the interaction mechanisms between agents to reduce misunderstandings and enhance collaboration efficiency.

References

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 439–451. Springer.

T Guo, X Chen, Y Wang, R Chang, S Pei, NV Chawla, O Wiest, and X Zhang. 2024a. Large language model based multi-agents: A survey of progress and challenges. In *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*. IJCAI; Cornell arxiv.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024b. Lightrag: Simple and fast retrieval-augmented generation.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. Rag-qa arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374.

Wanwei He, Jiabao Wen, Lei Zhang, Hao Cheng, Bowen Qin, Yunshui Li, Feng Jiang, Junying Chen, Benyou Wang, and Min Yang. 2023. Hanfei-1.0. https://github.com/siat-nlp/HanFei.

Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. 2022. Forming effective human-ai teams: Building machine learning models that complement the capabilities of multiple experts. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, page 2478.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Bo Li, Shuang Fan, and Jin Huang. 2024. Csaft: Continuous semantic augmentation fine-tuning for legal large language models. In *International Conference on Artificial Neural Networks*, pages 293–307. Springer.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

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

Zhen Wan, Yating Zhang, Yexiang Wang, Fei Cheng, and Sadao Kurohashi. 2024. Reformulating domain adaptation of large language models as adapt-retrieve-revise: A case study on chinese legal domain. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5030–5041.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Lei Wang, Jingsen Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Recagent: A novel simulation paradigm for recommender systems. *arXiv preprint arXiv:2306.02552*.

Yuhui Wang, Hao He, and Xiaoyang Tan. 2020. Truly proximal policy optimization. In *Uncertainty in artificial intelligence*, pages 113–122. PMLR.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*.

Shiguang Wu, Zhongkun Liu, Zhen Zhang, Zheng Chen, Wentao Deng, Wenhao Zhang, Jiyuan Yang, Zhitao Yao, Yougang Lyu, Xin Xin, Shen Gao, Pengjie Ren, Zhaochun Ren, and Zhumin Chen. 2023. fuzi.mingcha. https://github.com/irlab-sdu/fuzi.mingcha.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See Kiong Ng, and Jiashi Feng. 2024. Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7315–7332.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024a. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Xiaoxian Yang, Zhifeng Wang, Qi Wang, Ke Wei, Kaiqi Zhang, and Jiangang Shi. 2024b. Large language models for automated q&a involving legal documents: a survey on algorithms, frameworks and applications. *International Journal of Web Information Systems*, 20(4):413–435.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. Leven: A large-scale chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 183–201.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *CoRR*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023a. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023b. Huatuogpt, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 4435–4439.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *arXiv preprint arXiv:2406.04614*.

A A sample from Legal Consultation Report Generation

Figure 9 is a sample of legal consultation summary report. A summary report comprises nine sections: report number, consultation date, client, subject of consultation, purpose of consultation, facts and background, legal analysis, legal advice, and risk warnings.

报告编号<Report Number>：LC20240704

咨询日期<Consultation Date>：2024年7月4日 (July 4, 2024)

委托人<Client>：张三 (Zhang)

咨询事项< Consultation Issues >：关于劳动合同解除的法律咨询 (Legal Consultation on the Termination of Employment Contracts)

咨询目的< Purpose of Consultation >：了解劳动合同解除的合法性及可能产生的法律后果 (Understand the legality of labor contract termination and the potential legal consequences.)

事实和背景<Facts and Background>:
张三先生于2023年1月1日与XX科技有限公司签订了一份为期三年的固定期限劳动合同。合同中明确了张三先生的职位为软件工程师，月薪为15000元，并约定了工作地点、工作时间、休息休假、社会保险和公积金缴纳等相关条款。2024年5月，公司因业务调整，决定裁减部分员工，张三先生被列入裁员名单。公司提出，将按照法定标准支付一个月工资作为经济补偿，并要求张三先生在一周内完成工作交接并离职。张三先生对此表示异议，认为公司的解除合同程序不符合法律规定，并担心自身权益受损。(Mr. Zhang signed a three-year fixed-term labor contract with XX Technology Co., Ltd. on January 1, 2023. The contract specified that Mr. Zhang's position was software engineer, with a monthly salary of 15,000 yuan, and outlined details regarding the work location, working hours, rest and vacation, social insurance, and housing fund contributions. In May 2024, due to business adjustments, the company decided to lay off some employees, and Mr. Zhang was included in the layoff list. The company proposed to pay one month's salary as severance compensation according to legal standards and requested that Mr. Zhang complete the work handover and leave within a week. Mr. Zhang objected to this, believing that the company's contract termination procedures did not comply with legal regulations and was concerned about the potential infringement of his rights.)

法律分析<Legal Analysis>:
根据《中华人民共和国劳动合同法》第四十一条的规定，用人单位因生产经营需要裁减人员二十人以上或者裁减不足二十人但占企业职工总数百分之十以上的，应当提前三十日向工会或者全体职工说明情况，听取工会或者职工的意见后，裁减人员方案经向劳动行政部门报告，可以裁减人员。XX科技有限公司在未提前三十日向张三先生说明情况，也未听取其意见的情况下，直接提出解除劳动合同，程序上存在瑕疵。(According to Article 41 of the Labor Contract Law of the People's Republic of China, if an employer needs to lay off more than twenty employees or fewer than twenty employees but more than ten percent of the total number of employees due to production and operational needs, it shall explain the situation to the trade union or all employees thirty days in advance. After listening to the opinions of the trade union or the employees and reporting the layoff plan to the labor administrative department, the employer may proceed with the layoffs. XX Technology Co., Ltd. directly proposed to terminate the labor contract with Mr. Zhang without explaining the situation to him thirty days in advance or listening to his opinions, which is procedurally flawed.)
同时，根据《劳动合同法》第四十七条的规定，用人单位解除劳动合同应当按照员工在本单位工作的年限，每满一年支付相当于一个月工资的经济补偿。张三先生在公司工作已超过一年，公司仅提出支付一个月工资作为补偿，可能不符合法律规定的补偿标准。(Additionally, according to Article 47 of the Labor Contract Law, when an employer terminates a labor contract, it shall pay economic compensation equivalent to one month's salary for each full year of service with the employer. Mr. Zhang has worked for the company for more than one year, and the company only proposed to pay one month's salary as compensation, which may not meet the legal compensation standards.)

法律建议< Legal Advice >:
1. 协商解决：建议张三先生首先与公司进行协商，明确解除合同的具体原因，争取合理的经济补偿。在协商过程中，张三先生可以提出公司解除合同程序的不当之处，并要求公司依法支付相应的经济补偿。
2. 收集证据：在协商过程中，张三先生应注意收集与劳动合同解除相关的证据，包括但不限于劳动合同、工资单、公司的通知函、工作交接记录等，以备不时之需。
3. 法律途径：如果协商未能达成一致，张三先生可以向劳动仲裁委员会提起仲裁申请，要求确认公司解除合同的违法性，并请求支付合法的经济补偿。
4. 法律援助：考虑到劳动争议案件的专业性，张三先生可以寻求专业律师的帮助，以便更有效地维护自身权益。

风险提示< Risk Warning >:
1. 时间限制：劳动争议案件通常有时效性要求，张三先生应在知道或应当知道权利被侵害之日起一年内提起仲裁或诉讼，否则可能丧失法律保护的机会。
2. 证据保全：在劳动争议中，证据的完整性和有效性至关重要。张三先生应确保所有证据的保存，并在必要时进行公证，以增强证据的法律效力。
3. 经济补偿不足：如果公司支付的经济补偿低于法定标准，张三先生可能面临经济上的损失。

Figure 9: A sample consulting report for in-context learning

B Case Analysis Presentation

We present the responses of LawLuo and baselines to users' vague legal queries in Table 2.

C Prompt Template for Guiding GPT-4o in Scoring

Figure 10 is a prompt template for guiding GPT-4o in scoring.

请你作为一个公正的判别器，评估模型对以下用户问题的回答质量。你的评价应该考虑律师风格语言、法律建议的实用性、法律知识的准确性三个维度。评价要尽可能客观。提供解释性说明后，你必须严格遵循以下格式，在1到10的范围内对响应进行评分，例如[分数=5].
Please serve as an impartial evaluator to assess the quality of the model's response to the following user query. Your evaluation should consider three dimensions: lawyer-like language, usefulness of legal advice, and accuracy of legal knowledge. The evaluation must be as objective as possible. After providing explanatory notes, you must adhere strictly to the format below and rate the response on a scale of 1 to 10, for example, [score=5].

Figure 10: Prompt template for guiding GPT-4o in scoring

	Inquiry: 我的朋友借了我一些钱，现在他说他没有能力还款，我该怎么办？(My friend borrowed some money from me, and now they say they are unable to repay. What should I do?)
LawLuo	Response: 为了更好地帮助您解决问题，能否提供更多细节？比如借款金额、还款期限，是否有书面协议或相关证据（如转账记录），朋友的财务状况，以及您与朋友之间的沟通情况（是否讨论过分期还款等方案）。这些信息有助于我们更准确地分析您的问题并提出合适的建议。(To better assist you in resolving the issue, could you provide more details? For example, the loan amount, repayment term, whether there is a written agreement or related evidence (such as transfer records), your friend's financial situation, and the communication between you and your friend (whether installment repayment options were discussed, etc.). This information will help us analyze your situation more accurately and provide appropriate advice.)
ChatGLM-3	Response: 您可以和朋友协商解决，比如分期还款或延期。若协商不成，您可以通过法律途径，提起诉讼。(You can negotiate with your friend to find a solution, such as agreeing on installment payments or a delay in repayment. If negotiation fails, you can pursue legal action and file a lawsuit.)
BaiChuan	Response: 遇到朋友无法还款的情况，可以先与对方沟通，了解其实际经济状况，商定双方都可以接受的还款方式。如果协商未果，建议您咨询当地法律顾问，了解您所在地区关于债务纠纷的具体法律规定。(If you encounter a situation where your friend is unable to repay the loan, you can first communicate with them to understand their actual financial situation and agree on a repayment method that both parties can accept. If negotiation fails, it is advisable to consult a local legal advisor to understand the specific legal regulations regarding debt disputes in your area.)
LawGPT	Response: 建议首先与朋友保持良好沟通，尽量避免矛盾升级。您可以考虑灵活处理，比如延期或分期还款。如果问题无法解决，可以查阅相关法律条文或寻求律师帮助，以确保自己的合法权益得到保障。(It is advisable to maintain good communication with your friend first, in order to avoid escalating conflicts. You can consider flexible solutions, such as delaying or agreeing on installment payments. If the issue cannot be resolved, you can consult relevant legal provisions or seek help from a lawyer to ensure your legal rights are protected.)
HanFei	Response: 此类问题的处理需要结合具体情况分析。建议您先尝试与朋友协商，充分了解其还款困难的原因，或制定双方认可的分期还款计划。如果协商无果，可以准备相关证据（如借条、转账记录等），通过正式的法律途径解决。同时，在追求法律手段前，需评估其执行能力和成本效益。(The resolution of such issues requires analyzing the specific situation. It is recommended to first try negotiating with your friend, understanding the reasons behind their repayment difficulties, or creating an installment plan both parties agree on. If negotiations fail, you can gather relevant evidence (such as loan receipts or transfer records) and pursue legal action. However, before opting for legal means, it is important to evaluate the enforceability and cost-effectiveness of this approach.)

Table 2: Responses to Inquiry about Loan Repayment Issue