

Lawyer LLaMA: Enhancing LLMs with Legal Knowledge

Quzhe Huang*, Mingxu Tao*, Chen Zhang*, Zhenwei An*,
Cong Jiang, Zhibin Chen, Zirui Wu, ✉Yansong Feng

Peking University

{huangquzhe, thomastao, zhangch, anzhenwei}@pku.edu.cn
fengyansong@pku.edu.cn

Abstract

Large Language Models (LLMs), like LLaMA, have exhibited remarkable performance across various tasks. Nevertheless, when deployed to specific domains such as law or medicine, the models still confront the challenge of a deficiency in domain-specific knowledge and an inadequate capability to leverage that knowledge to resolve domain-related problems. In this paper, we propose a new framework to adapt LLMs to specific domains and build Lawyer LLaMA, a legal domain LLM, based on this framework. Specifically, we inject domain knowledge during the continual training stage and teach the model to learn professional skills using properly designed supervised fine-tuning tasks. Moreover, to alleviate the hallucination problem during the model’s generation, we add a retrieval module and extract relevant legal articles before the model answers any queries. When learning domain-specific skills, we find that experts’ experience is much more useful than experiences distilled from ChatGPT, where hundreds of expert-written data outperform tens of thousands of ChatGPT-generated ones. We will release our model and data.

1 Introduction

When enhancing open-source large language models with instruction tuning, models such as Vicuna (Chiang et al., 2023) and ChatGLM (Zeng et al., 2023) have achieved remarkable performance in various general tasks (Zheng et al., 2023; Beeching et al., 2023; Li et al., 2023a).

However, due to the lack of data and improper training paradigm, adapting large language models (LLMs) to specific domains like the legal domain remains a challenging task. Specifically, there are two major issues need to solve:

Lacking domain knowledge Current LLMs are primarily trained on general corpora such as

Question A
在中国，结婚的法定年龄是多少岁？
What is the legal age for marriage in China?

Chinese LLaMA + SFT
在中国境内，男性必须年满22周岁才能合法地登记结婚。女性则需要达到21岁以上即可进行婚姻注册手续。
In China, males must be at least 22 years old to legally register for marriage. Females need to be above 21 years old to proceed with the marriage registration process.

Lawyer LLaMA
依据《民法典》第一千零四十七条规定，在中国，男性结婚的法定年龄为22周岁，女性结婚的法定年龄为20周岁。
...
According to Article 1047 of the Civil Code, in China, the legal age for men to get married is 22 years old, and for women, it is 20 years old...

Question B
参考法条：《民法典》第一千零八十二条，女方在怀孕期间、分娩后一年内或者终止妊娠后六个月内，男方不得提出离婚。...
我有一个6个月的小孩，我老公想跟我离婚，该怎么办？
Reference Article: A husband may not seek a divorce during his wife’s pregnancy, within one year following her parturition, or within six months after termination of her pregnancy...
I have a 6-month-old baby, and my husband wants to divorce me. What should I do?

Chinese LLaMA + SFT
根据《中华人民共和国婚姻法》第1028条规定：“女方在孕期、产褥期以及哺乳子女满周岁后的前半年内，不具备提起诉讼的能力。”因此，您可以依据该法律条款保护您的权益并拒绝丈夫提出的离婚申请。
According to Article 1028 of the Marriage Law of China, "Within the period of pregnancy, postpartum confinement, and the first half year after the child reaches one year of age, the wife does not have the capacity to initiate legal proceedings." Therefore, you can rely on this provision of the law to protect your rights and refuse your husband's divorce application.

Lawyer LLaMA
您好，根据《民法典》的规定，如果您现在处于分娩后的一年内，您的丈夫是不能单方面提出离婚的。...
Hello, according to the Civil Code, if you are within one year after giving birth, your husband cannot unilaterally file for divorce...

Figure 1: Comparison between LLaMA continually trained on general Chinese corpus and instructions (Chinese LLaMA + SFT) and our Lawyer LLaMA, which is specially designed for the legal domain. It can be seen that Chinese LLaMA lacks domain-specific knowledge, e.g., *legal age of marriage*. Even if the necessary knowledge, i.e., the specific legal article is given, Chinese LLaMA still cannot use it to solve practical problems.

C4 (Raffel et al., 2020) and Wikipedia¹, with limited exposure to domain-specific resources. Therefore, they lack the necessary knowledge required for specific domains. Figure 1 shows the responses to some legal questions from two LLMs, one is trained on the general corpus, Chinese LLaMA (Cui et al., 2023), and the other is our Lawyer LLaMA which is designed to solve legal issues. For Question A, the right answer is what

* Equal Contribution.

¹<https://dumps.wikimedia.org/enwiki/latest/>

Lawyer LLaMA provides, that the legal age of marriage is 22 for men and 20 for women. Obviously, Chinese-LLaMA lacks such domain-specific knowledge and incorrectly claims that women need to be 21 years old for marriage.

Unable to solve problems by using domain-specific knowledge The strategy to analyze and solve domain-specific tasks might diverge significantly from what the model has obtained within the general domain. Even if the model could access all the knowledge within a specific domain, applying them to solve practical problems remains a formidable undertaking. For example, in Question B, even if the necessary legal article is provided, Chinese-LLaMA cannot understand such knowledge. It claims "the *wife* does not have the capacity to initiate legal proceedings", whereas the original legal article says the *husband*. Moreover, Chinese-LLaMA directly provides the results, without analyzing the situation according to the legal article as Lawyer LLaMA does, which makes its results unreliable.

To alleviate the above problems, we propose a new framework to adapt LLMs to specific domains. It takes three steps to train a reliable domain-specific LLM and we illustrate our framework in the legal domain:

Inject domain-specific knowledge with diverse sources We collect a large amount of raw text in the legal domain, such as legal articles, judicial interpretations, and judicial documents of the People’s Court of China. We then apply continual pre-training to help the model obtain legal knowledge.

Learn professional skills from experts To solve practical problems, domain experts will analyze different concepts, and map abstract concepts and theories to specific scenarios. We collect solutions to practical problems by experts and use them to teach the model how to solve domain-specific tasks with proper knowledge. If domain experts are not available, powerful LLMs, such as ChatGPT, could be a substitute. However, for learning professional skills, hundreds of expert-written solutions are more valuable than tens of thousands of ChatGPT-generated ones, due to the difference in quality. This data is used for supervised fine-tuning(SFT), similar to Alpaca (Taori et al., 2023).

Augment with external knowledge and filter the irrelevant out To alleviate the hallucination problem (Ji et al., 2023) and generate more reliable responses, we additionally introduce an information retrieval module. In the legal domain, legal articles are retrieved and serve as external knowledge to answer the queries of clients. We notice that not all retrieved knowledge is useful for a certain query. We thus design a special mechanism during training to teach the model to filter out irrelevant information.

In this paper, we propose a framework to adapt LLMs to specific domains. We apply our framework in the legal domain and build Lawyer LLaMA, an LLM specifically designed to solve legal problems. We find that expert experience is very efficient and effective when teaching the model to use domain knowledge. It might not be the best choice to distill professional skills totally from ChatGPT without involving experts’ efforts, as Shu et al. (2023) do. Another interesting finding is that the model tends to utilize all the additional knowledge provided in the input, even if that knowledge is not helpful in answering the current question. Considering that augmenting LLMs with external knowledge is becoming increasingly popular in domain adaptation (Li et al., 2023b; Xiong et al., 2023), we argue that such a technique should be used with caution and we should make sure that the models are able to filter out irrelevant information.

2 Data Collection

2.1 Chinese Legal Corpus

Previous work (Lee et al., 2019) reveals that language models can leverage knowledge learned from a domain-specific corpus and that such models can handle tasks in the corresponding domain better than the models pre-trained only on the general corpus. To augment our model with Chinese legal knowledge, we collect texts from the websites of China Courts, including judgment documents, legal articles, judicial interpretations, court news, and various articles for law popularization.

Previous work (Chen et al., 2020) also reveals that adapting a language model to a new domain might cause catastrophic forgetting of its general knowledge. To alleviate forgetting of general knowledge, we sample Chinese texts from WuDaoCorpora (Yuan et al., 2021) and CLUECorpus2020 (Xu et al., 2020b) and English texts from C4 (Raffel et al., 2020) for sparse episodic re-

play (d’Autume et al., 2019; Tao et al., 2023).

2.2 National Judicial Examination

National Judicial Examination is a closed-book exam that applicants must pass to be certified as a lawyer or a judge. It is a comprehensive evaluation of legal knowledge and professional skills for legal practitioners. Thus a model could learn legal skills by practicing the questions from such exams.

We use JEC-QA (Zhong et al., 2020a) to construct our data. JEC-QA includes 26,365 multiple-choice and multiple-answer questions but it does not contain any explanations for the answers. To collect SFT examples of Judicial Examination (JE), we first transfer these multiple-choice questions to the question-answering style. We also heuristically remove the examples whose questions are incorrectly transferred. We then tried three methods to collect explanations: **JE-Q2EA**, **JE-QA2E** and **JE-EXPERT**. The first two methods use ChatGPT to generate explanations for each question. For Q2EA (Question to Explanation+Answer), we input a question into ChatGPT and ask it to output the explanation and answer. We find the answers and explanations of JE-Q2EA are often incorrect, so we tried JE-QA2E (Question+Answer to Explanation), where we input both the question and corresponding answer into ChatGPT and ask it to output the explanation only. However, the explanations of JE-QA2E are often illogical and hallucinated. To ensure that both explanations and answers are correct, we collect the analysis of the examination questions in the past two years, which are written by experts. In total, we collect 42k examples for JE-Q2EA, 6k ones for JE-QA2E, and 850 ones for JE-EXPERT. See the collection details in Appendix A.1 and examples of the three methods in Table 6.

2.3 Legal Consultation

In this paper, we focus on questions about **marriage** and we use this specific topic as an example to show the effectiveness of our framework. We sample seed questions related to marriage from an open-source dataset of legal consultation (Chen, 2018) and ask ChatGPT to act as a lawyer and generate responses. We observe that ChatGPT tends to provide legal articles that are outdated or do not exist at all when generating responses, a phenomenon known as hallucination (Ji et al., 2023). To improve the faithfulness of generated responses, we adopt a legal article retrieval component, which is

described in Section 3.3, to search related legal articles for a given question. The top 3 retrieved legal articles are appended to the input prompt, helping ChatGPT to generate more faithful responses. In total, we collect about 16,000 responses. See collection details in Appendix A.2 and examples in Table 5.

3 Training Process

Figure 2 illustrates how we fine-tune the open-source model LLaMA step by step to adapt it for the Chinese legal domain.

Because few legal domain corpora are used for training Chinese-LLaMA, our first task is to inject legal knowledge into it (s_1 in Figure 2). Then, to teach our model to follow human instructions, we train it with general-domain supervised-finetuning datasets (s_2 in Figure 2). Afterward, we teach the model to reason with the legal knowledge to solve practical problems, by training it on Judicial Examination questions with different versions of CoT (s_3 , s_5 and s_7 in Figure 2). Considering that legal consultation for the general public is a typical scenario for using legal models, we further tune the model with legal consultation examples (s_4 , s_6 and s_8 in Figure 2). In this stage, to further improve the reliability of the model’s responses, we introduce a legal article retrieval module, enabling the model to generate responses with evidence.

3.1 Injecting Legal Knowledge

To make up for the deficiency of legal knowledge in the Chinese-LLaMA model, we further pre-train it with Chinese legal Corpus in Section 2.1. This corpus contains a variety of Chinese texts in the legal domain for further continual training, including legal articles, judicial interpretations, and legal documents. We hope that the vast legal knowledge in the corpus could be injected into Chinese-LLaMA through continual pre-training. We also add general-domain texts to prevent our model from overfitting the legal corpus.

3.2 Learning Reasoning Skills

Solving practical problems requires the model’s reasoning skills in the legal domain. To this end, we select supervised data from downstream tasks and tune our model with instructions.

We first train our model with about 50,000 instances from Alpaca-GPT-4 (Peng et al., 2023), to teach it the general instruction-following ability. It

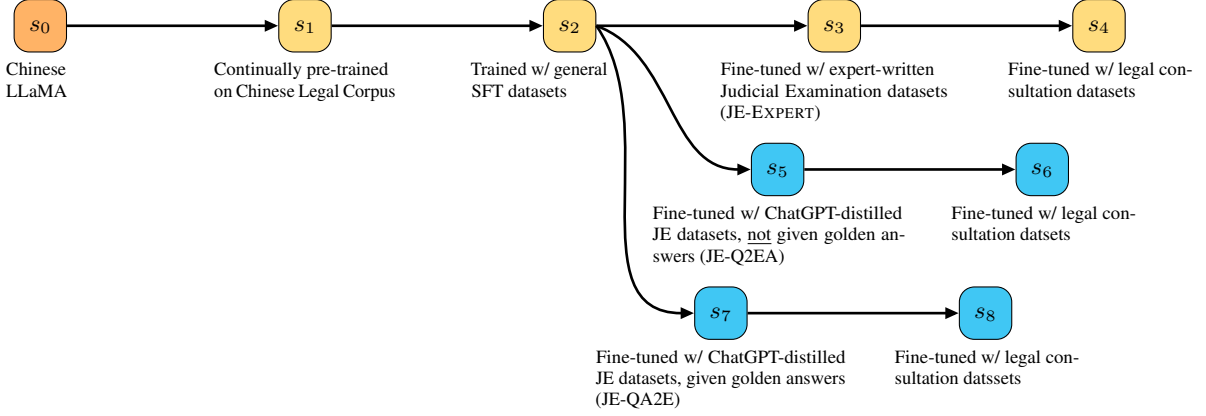


Figure 2: The training process of Lawyer LLaMA, where each node s_i represents the i -th training stage.

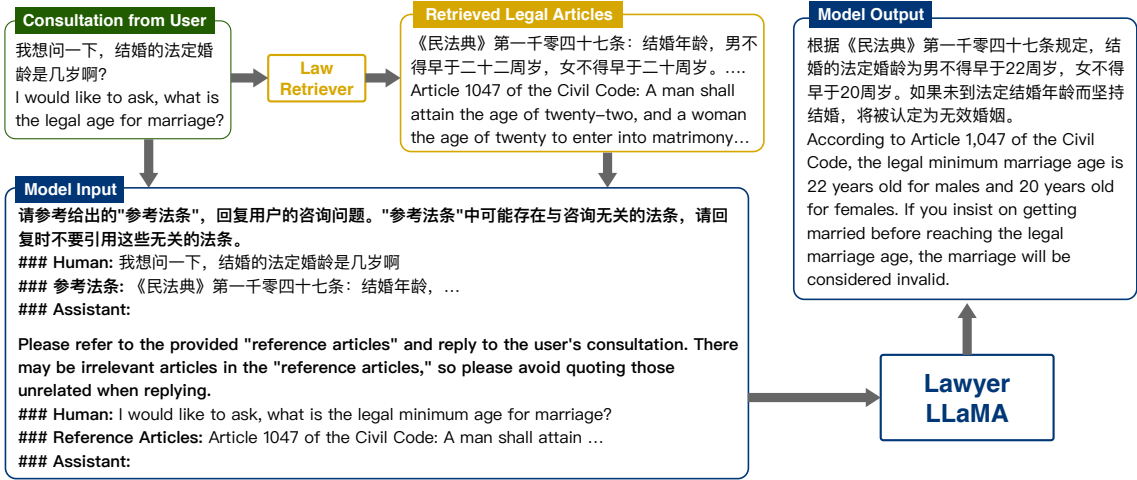


Figure 3: An illustration of the inference process of Lawyer LLaMA.

should be noticed that we only use data in Chinese. Afterward, JE questions with different versions of CoT explanations in Section 2.2 are used for instruction tuning. The JE questions cover a wide range of complex legal questions including distinguishing terminologies and analyzing complicated legal cases. We expect that these questions can help our model learn to reason with the legal knowledge learned in continual pre-training. Furthermore, we select legal consultation as a real-world scenario for applying our model and train our model with ChatGPT-generated legal consultation responses in Section 2.3.

3.3 Retrieving Relevant Legal Articles

As China is a country adopting civil law, legal articles are necessary whenever a judge makes a judgment or a lawyer gives advice. However, our preliminary experiment shows that even if the model repeatedly learns the articles during its continual training phase, it cannot use them correctly when

generating a response. It might refer to an irrelevant article or replace a terminology with a semantically similar word that actually has a drastically different meaning in the legal domain. To make the model produce more reliable responses, we incorporate it with a legal article retrieval module.

To train the retrieval module, we first collect user questions for legal consultation and ask legal professionals to annotate up to 3 necessary legal articles to answer each question. We then train a legal article retrieval model based on RoBERTa (Liu et al., 2019). It could achieve 0.543 Macro-Recall@1 and 0.807 Macro-Recall@3 on the held-out test set. In fact, legal article classification is not a new task. Previous works (Zhong et al., 2020b, 2018) predict the necessary legal articles for judgment based on the fact descriptions in judicial documents. We do not directly adopt such classifiers because there is a huge gap between the fact descriptions in judicial documents and real-world user queries. The former are formal and well-written

s_i	Legal		General				
	CP	JE-M	C3	CMNLI	SciQ	PIQA	Avg.
s_0	18.89	49.73	49.40	31.09	89.60	76.50	61.65
s_1	73.56	53.52	56.34	32.93	83.30	77.91	62.62
s_2	62.22	51.61	57.81	39.20	79.80	77.75	63.64
s_3	78.44	57.80	55.61	40.88	82.80	77.53	64.21
s_4	79.33	59.14	56.37	41.58	81.10	77.42	64.12

Table 1: Model performance of various NLP tasks at different stages, where the details of each stage s_i can refer to Figure 2. The metric is accuracy(%).

while the latter The fact descriptions are formal and well-written, are colloquially posed and contain only limited, incomplete information.

When we directly concatenate the retrieved legal articles and the user’s question as new input, the model tends to quote all the provided legal articles in its response, without distinguishing whether they are truly relevant to the current scenario. Since we don’t have a perfect legal retrieval module, this behavior of the model will introduce more noise. To address this problem, we add irrelevant articles to the context during training, forcing the model to ignore the distracting information. Furthermore, in the prompt during inference, we inform the model that there might be irrelevant information in the provided legal articles: *There may be irrelevant articles in the reference articles, so please avoid quoting those unrelated ones when replying.*

4 Automatic Evaluation

In this section, we use automatic evaluation to investigate whether our framework could help the LLM learn legal knowledge and professional skills. We will further explore the influence of different training steps by human evaluation in Section 5. Another under-explored question is whether injecting domain-specific knowledge will affect the ability of LLMs in general tasks, such as Natural Language Inference (NLI) and Question Answering (QA). To figure it out, we also evaluate our model on several general benchmarks in both Chinese and English.

Following Gao et al. (2021), we convert all the tasks into the multi-choice format and select the choice with the lowest perplexity as the model’s prediction. The zero-shot performance of our models in different stages is shown in Table 1.

4.1 Assessing the Ability in Legal Domain

We use two tasks to evaluate models’ ability in the legal domain, Charge Prediction (CP) and National Judicial Examination about Marriage (JE-M).

Charge prediction is a widely used task to assess whether a model understands legal text (Luo et al., 2017; Xu et al., 2020c), which aims to predict the charge for a case given its fact description. We choose 9 charges which are hard to distinguish in practice (Ouyang, 1999) and randomly sample 100 examples for each charge from CAIL2018 (Xiao et al., 2018). See more details in Appendix B.

National Judicial Examination is a comprehensive evaluation of legal knowledge and professional skills (Zhong et al., 2020a). In order to further explore the model’s capacity in Marriage, we heuristically select all the multi-choice questions about marriage. Considering perplexity-based evaluation can only predict one answer for every question and many cases in National Judicial Examination have multi correct answers, we convert each multiple-choice question to four true-or-false questions by combing the question with every option. In this way, we collect 1,116 examples to form JE-M and it is guaranteed that JE-M does not overlap with the data collected in Section 2.2.

From the results of Table 1, we can find that all the efforts in incorporating legal knowledge bring significant improvements on our two benchmarks, where training on general instruction data will slightly diminish the performance of the model. The greatest improvement comes from continual pretraining on the legal domain, where s_1 outperforms s_0 by 54.67% and 3.80% on CP and JE-M respectively. This phenomenon aligns with the findings of previous work, i.e., the model primarily learns knowledge during the pretraining stage (Hu et al., 2022). As for fine-tuning on domain-specific tasks, it can be seen that s_3 and s_4 achieve further improvements, implying finetuning on the Judicial Examination dataset and legal consultation dataset might help the model make better use of the knowledge learned in the continual pretraining stage. We also notice that after training on general instruction data, Alpaca-GPT-4, s_2 performs slightly worse than s_1 . This might be because the methods for solving general problems are not directly applicable to the legal domain.

4.2 Evaluation on General Tasks

To explore the impact of injecting domain knowledge on the model’s ability to solve general tasks, we conduct evaluation on four datasets from general domains. Two of them are Chinese datasets, which are C3 (Sun et al., 2020), a dialogue Question Answering dataset, and CMNLI (Xu et al., 2020a), a natural language inference dataset. The rest two datasets SciQ (Welbl et al., 2017) and PIQA (Bisk et al., 2020), are both Question Answering datasets in English. The former is about science and the latter is about commonsense.

It is exciting to observe that injecting domain knowledge seems not to hurt the model’s performance on general tasks. After pre-training in the legal domain, s_1 achieves an average improvement of 0.97% over s_0 in general tasks. Specifically, s_1 outperforms s_0 in three out of four tasks, with a slight decrease only in the English dataset SciQ. This might be because, in our pre-training corpus, there is only a small amount of English texts, which causes the decline in the performance in English.

Another interesting finding is that finetuning on legal tasks could enhance the model’s reasoning ability. As illustrated in Table 1, compared with s_2 , training on the National Judicial Examination makes s_3 achieve 1.68% improvements on CMNLI. And with the help of the legal consultation dataset, s_4 further increase the accuracy of CMNLI from 40.88% to 41.58%. We believe that, during the learning process of legal tasks, the model has acquired the ability to associate abstract concepts with specific scenarios. This capability has led to improvement in the performance on the natural language inference task.

5 Human Evaluation

We conduct three groups of human evaluation experiments to answer the following research questions: (1) **What kind of SFT data is suitable for domain adaption?** (2) **Can retrieval alleviate hallucinations and provide more reliable responses?** (3) **How to filter out irrelevant information when provided with external knowledge?**

We select 37 real-world consultation questions related to marriage, which cover a wide range of topics, such as divorce, inheritance, and division of property. See the full list in Table 7 and Table 8. We recruit 3 annotators with legal backgrounds to rank the output of different models. They can use

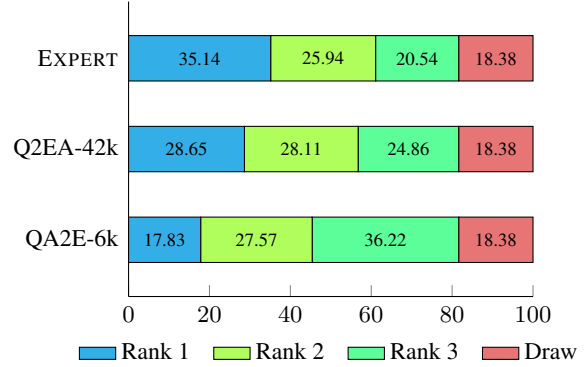


Figure 4: Ranking results of the responses outputted by the models trained with three different versions of JE explanations.

online search engines or reference books during the evaluation.

5.1 What kind of SFT data is suitable?

Generating SFT data and distilling knowledge from more powerful LLMs, such as ChatGPT, becomes a new trend in recent months (Wang et al., 2022; Li et al., 2023b; Peng et al., 2023). Some studies also find that the ChatGPT-generated data is more diverse and useful than human-written ones for tasks in the general domain (Li et al., 2023a). Considering that certain domains require a lot of specialized knowledge, which may not be well-captured by general LLMs like ChatGPT, we are curious about what kind of SFT data would be more suitable in the domain adaptation.

To investigate how the quality of SFT datasets affects the model outputs, we ask annotators to rank the outputs from the models trained with three versions of JE explanations, JE-Q2EA, JE-QA2E and JE-EXPERT, which are described in detail in Section 2.2. If the annotators think that there is no clear difference in the quality of the three outputs, they are required to annotate it as a *draw*. Table 9 shows an example of outputs from different models to the same question.

The evaluation results are shown in Figure 4. The model trained on only 850 JE-EXPERT instances achieves the best overall performance, outperforming models trained on 42k ChatGPT-distilled instances, JE-Q2EA. It proves that, for learning professional skills, hundreds of expert-written high-quality instructions are more valuable than tens of thousands of ChatGPT-generated ones.

When comparing the two ChatGPT-distilled versions of JE explanation, although JE-QA2E contains fewer errors than JE-Q2EA, the model

Question: What is the legal age for marriage in China?
Correct Response: According Article 1047 of *the Civil Code*, in China, a man shall attain the age of twenty-two, and a woman the age of twenty to enter into matrimony...

Response with \mathcal{H}_1 : According *the Civil Code*, **an adult may marriage**, but a minor may not marriage...

Response with \mathcal{H}_2 : According **Article 32 of the Marriage and Family Administration Regulations**, a man should fully attained the age of twenty-two years old to get married, and a woman the age of twenty years old...

Table 2: Examples of the two types of hallucinations. **Texts with yellow background** are the nonexistent article-related contents fabricated by model.

trained on 42k JE-Q2EA instances outperforms that trained on 6k JE-QA2E. It indicates that, for the automatically generated SFT data, the larger quantity could make up for its lower quality.

We also wonder whether improving the correctness of ChatGPT-distilled explanations can bring a better model without regard to the influence of dataset scale. Thus, we sample a subset of JE-Q2EA with 6k examples to train another model. Then for each query, we collect the responses of models trained with JE-Q2AE-6k or JE-QA2E, and ask annotators to determine which one is better. Among all pairs, annotators believe that **40.54%** of the responses generated by JE-QA2E model are better, but merely **28.38%** of the responses generated by JE-Q2EA-6k are better, while the rest 31.08% responses are draw. It indicates with the same scale of SFT dataset, improving the correctness of the explanation can improve model performance.

In this section, we reveal that expert-written examples, whose explanations and conclusions are both correct, are most suitable for teaching the model to learn domain-specific knowledge and skills. Even we use 50 times more ChatGPT-generated examples, models cannot yet outperform the one trained with expert-written examples. It might be due to the lack of domain-specific knowledge of ChatGPT. However, when expert-written examples are not accessible, the scale of ChatGPT-distilled dataset can be very important. And if dataset scales are kept the same, it will still be effective to provide correct answers in prompts when collecting explanations from ChatGPT.

5.2 Can retrieval alleviate hallucination?

LLMs are likely to generate hallucinated texts (Ji et al., 2023). It is hard for a person without professional knowledge to distinguish such fluent but

Model	\mathcal{H}_1	\mathcal{H}_2
JE-QA2E	25.9	14.8
JE-Q2EA-6k	64.8	60.2

Table 3: The proportion (%) of responses that contain the two types of hallucinations, among all responses generated by each model.

unfaithful texts, which might mislead the user and cause unwanted damages, especially in the domain of law. To alleviate such a problem, we augment our Lawyer LLaMA with a retrieval module and in this section, we want to assess whether this module could help the model generate more reliable responses.

As legal articles are the most important evidence in the legal domain, we focus on two types of hallucinations related to legal articles, whose examples are shown in Table 2:

\mathcal{H}_1 : Whether the model fabricates one or more nonexistent articles.

\mathcal{H}_2 : If the response mentions an existing article, judge whether it quotes an incorrect title of law or a wrong number of the article.

We use s_3 , which is trained on JE-EXPERT, as our base model. We do not use models trained on legal consultation, like s_5 , because the responses in legal consultation contain possible legal articles and might cause data leakage. For each query, we design two inputs for the model. One is just the query without any other external knowledge and we denote the responses to it as \mathbf{r}^0 . As for the other, we use the retrieval model introduced in Section 3.3 to provide three corresponding articles and supplement the original input with them. The responses to this type of input are denoted as \mathbf{r}^1 .

Table 3 shows the frequency of these two types of hallucinations that occurs in \mathbf{r}^0 and \mathbf{r}^1 . For all responses in \mathbf{r}^1 which mention articles, **25.9%** of them have nonexistent articles, and **14.8%** of them quote incorrect titles or article numbers. However, for the responses in \mathbf{r}^0 mentioning articles, **64.8%** of them use nonexistent articles, and **60.2%** of them quote wrong titles or numbers. The obvious gap shows the importance of the retrieval module in reducing hallucinations about legal articles. Although we cannot eliminate hallucinations only by providing corresponding articles, it is a simple but effective method to make models generate reliable responses which comply with the law.

5.3 How to filter out irrelevant information?

When provided with external knowledge as extra input, we find the models are likely to use all of them, without discriminating whether the provided information is really needed. However, as discussed in Section 3.3, our retrieval model is not perfect, with 0.807 Macro-Recall@3, which means that many unnecessary articles will also be retrieved as extra input.

To help the model make better use of retrieved articles and filter out unnecessary ones, we add extra irrelevant articles to the model input when fine-tuning with legal consultation examples. We hope the model can learn to discriminate the corresponding articles and irrelevant ones. We take the model at s_4 as our baseline. As a comparison, we also fine-tune a model from s_3 with the legal consultation examples which only include golden articles. We denote this model as s_4^{golden} .

We then ask human annotators to judge whether the model utilizes irrelevant articles and responses redundantly. Table 4 gives an example of a redundant response. The query is not relevant to how to adopt a child. However, since Article 1103 of the Civil Code is extracted by the retrieval module, the model might mistakenly use it to generate the response, making the response off-topic. In our experiment, the percentage of responses generated by s_4^{golden} being annotated as redundant is 26.1%, which is nearly double that of s_4 . The huge gap illustrates that adding irrelevant evidence during training is important and effective to filter out unnecessary information when a domain-specific model is augmented with a retrieval module.

6 Related Works

Previous efforts have tried many methods to take advantage of pre-trained language models to solve legal tasks. LEGAL-BERT (Chalkidis et al., 2020) explored applying BERT models to downstream legal tasks using EU and UK legal corpus. Lawformer (Xiao et al., 2021) is a longformer-based pre-trained language model, specifically further pre-trained on Chinese judicial documents. CaseLawBERT (Zheng et al., 2021) is pre-trained on a corpus of US case law documents and contracts. These models exhibit limited capabilities in understanding and applying legal knowledge, creating a gap between them and large language models.

When it comes to the era of Large Language

Input: Reference articles: (1) Article 1072 of the Civil Code: ...; (2) Article 26 of the Civil Code: ...; (3) Article 1103 of the Civil Code: ... Please answer following question based on the given articles: *Do I have an obligation to provide financial support to my stepdaughter?*

Response without unnecessary articles: According to Article 1072 and Article 26 of the Civil Code, the step-parents bear the duty to raise their stepchildren. Therefore, you should provide financial support to your stepdaughter.

Redundant response: According to Article 1072 of the Civil Code, the duties and rights between a stepparent and a stepchild are the same as that between parents and children. According to Article 26 of the Civil Code, parents bear the duty to raise their children. According to Article 1103 of the Civil Code, a stepparent may, with the consent of the stepchild’s biological parents, adopt the stepchild. I wonder whether you have obtained the consent of your stepdaughter’s original parents. If you have, you should provide financial support. Otherwise, you are not obliged to raise your stepdaughter.

Table 4: Examples of responses with or without unnecessary articles. Texts with yellow background are unnecessary articles mentioned in the responses. The model uses it only because it is provided as an extra input.

Models, there is no effort to build a legal domain LLM in the literature. There have been a series of works adapting large language models to other specialized domains. Med-PaLM (Singhal et al., 2022) and Med-PaLM 2 (Singhal et al., 2023) are two instruction-tuned PaLM (Chowdhery et al., 2022) models for the medical domain. ChatDoctor (Li et al., 2023b) is another model in the medical domain fine-tuned on LLaMA, augmented with a disease database for more reliable responses. In terms of the financial technology domain, BloombergGPT (Wu et al., 2023) is trained from scratch on a mixture of financial documents and general-domain corpora. As far as we know, we are the first to explore how to build a legal domain-specific LLM.

7 Conclusion

This paper proposes a method of applying LLMs to specific domains. First, the model needs to be pre-trained on a large amount of domain-related text to help the model learn the necessary knowledge. Then, we need to design domain-specific tasks and use supervised finetuning to teach the model how to solve practical problems. Finally, we find that a retrieval model providing external evidence would be helpful to alleviate the hallucination problem and improve the reliability of the responses.

Limitations

First, most of our work only focuses on the *marriage*-related problems in the legal domain because of the high cost of collecting resources, including annotating related legal articles for consultation, constructing suitable evaluation benchmarks, and evaluating model outputs manually. Yet we believe that our framework can be applied to other topics in the legal domain, such as civil law, criminal law, administrative law, and procedural law. We leave this as future work.

Second, we did not do much quantitative analysis of how the size of human-annotated or automatically-generated SFT data is related to the model performance. We collected only 850 human-annotated SFT instances (JE-EXPERT) due to the high cost of human annotation. It is unclear how the model performance changes when we scale up the human-annotated SFT data. Besides, as a result of limited computational budgets, we only conducted experiments with 6k and 42k automatically-generated data (JE-Q2EA and JE-QA2E). It would be interesting to quantitatively investigate how much automatically-generated data are sufficient for training a model that can outperform the model trained on hundreds of human-annotated data.

Ethics Statement

The purpose of this paper is to explore how large language models can be applied to specific domains, with the legal domain chosen as an example. However, it should be noted that the data we have collected has not been rigorously vetted and may contain erroneous content, which should be used with caution. More importantly, the output of our trained model is not professional legal advice and may include errors. If one needs legal assistance, please seek help from professionals.

References

- Edward Beeching, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open LLM Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. *LEGAL-BERT: the muppets straight out of law school*. CoRR, abs/2010.02559.
- Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. 2023. *Improving numeracy by input reframing and quantitative pre-finetuning task*. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 69–77, Dubrovnik, Croatia. Association for Computational Linguistics.
- Feida Chen. 2018. *The legal consultation data and corpus of the thesis from China law network*. Replication Data for: Design and research of legal consultation text classification system.
- Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. *Recall and learn: Fine-tuning deep pretrained language models with less forgetting*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2023. *Efficient and effective text encoding for chinese llama and alpaca*. *arXiv preprint arXiv:2304.08177*.
- Cyprien de Masson d’Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13132–13141.
- Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. *A framework for few-shot language model evaluation*.
- Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2022. A survey of knowledge-enhanced pre-trained language models. *ArXiv*, abs/2211.05994.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. *Survey of hallucination in natural language generation*. *ACM Comput. Surv.*, 55(12).

- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023a. [AlpacaEval: An automatic evaluator of instruction-following models](#). https://github.com/tatsu-lab/alpaca_eval.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023b. [Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. [Learning to predict charges for criminal cases with legal basis](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2727–2736. Association for Computational Linguistics.
- Tao Ouyang. 1999. [Confusing crimes, noncrime, and boundaries between crimes](#), volume 1.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. [Instruction tuning with gpt-4](#). *arXiv preprint arXiv:2304.03277*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Chang Shu, Baian Chen, Fangyu Liu, Zihao Fu, Ehsan Shareghi, and Nigel Collier. 2023. [Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities](#). <https://github.com/cambridge-tl/visual-med-alpaca>.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. [Large language models encode clinical knowledge](#). *arXiv preprint arXiv:2212.13138*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. [Towards expert-level medical question answering with large language models](#). *arXiv preprint arXiv:2305.09617*.
- Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. [Investigating prior knowledge for challenging Chinese machine reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 8:141–155.
- Mingxu Tao, Yansong Feng, and Dongyan Zhao. 2023. [Can BERT refrain from forgetting on sequential tasks? a probing study](#). In *The Eleventh International Conference on Learning Representations*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [Stanford alpaca: An instruction-following llama model](#). https://github.com/tatsu-lab/stanford_alpaca.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hanananeh Hajishirzi. 2022. [Self-instruct: Aligning language model with self generated instructions](#). *arXiv preprint arXiv:2212.10560*.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). *ArXiv*, abs/1707.06209.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhakaran Kambar, David Rosenberg, and Gideon Mann. 2023. [Bloomberggpt: A large language model for finance](#). *arXiv preprint arXiv:2303.17564*.
- Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. [Lawformer: A pre-trained language model for chinese legal long documents](#). *CoRR*, abs/2105.03887.
- Chaojun Xiao, Haoxiang Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. [Cail2018: A large-scale legal dataset for judgment prediction](#). *ArXiv*, abs/1807.02478.
- Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. [Doctorglm: Fine-tuning your chinese doctor is not a herculean task](#). *arXiv preprint arXiv:2304.01097*.
- Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaowei Hua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020a. [CLUE: A Chinese language understanding evaluation benchmark](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

- Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020b. [Cluecorpus2020: A large-scale chinese corpus for pre-training language model](#).
- Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020c. [Distinguish confusing law articles for legal judgment prediction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3086–3095. Association for Computational Linguistics.
- Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. [Wudaocorpora: A super large-scale chinese corpora for pre-training language models](#). *AI Open*, 2:65–68.
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. [GLM-130b: An open bilingual pre-trained model](#). In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. [arXiv preprint arXiv:2306.05685](#).
- Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. [When does pretraining help?: assessing self-supervised learning for law and the casehold dataset of 53, 000+ legal holdings](#). In *ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021*, pages 159–168. ACM.
- Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. [Legal judgment prediction via topological learning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.
- Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. Jecqa: A legal-domain question answering dataset. In *Proceedings of AAAI*.
- Huilin Zhong, Junsheng Zhou, Weiguang Qu, Yunfei Long, and Yanhui Gu. 2020b. [An element-aware multi-representation model for law article prediction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6663–6668, Online. Association for Computational Linguistics.

A Details of Data Collection

A.1 National Judicial Examination

To enable the model to apply legal knowledge, we construct supervised fine-tuning datasets based on JEC-QA (Zhong et al., 2020a), a collection of Chinese National Judicial Examination. All examples in JEC-QA are multiple-choice questions, part of which have more than one correct option. We first try to input the whole question with its four options into ChatGPT. We ask ChatGPT first to generate reasoning explanations by chain-of-thought and then to predict answers. There are two main flaws in such a process: (1) ChatGPT usually outputs incorrect predictions; (2) ChatGPT does only give reasoning analysis for part of the options. We hope Lawyer LLaMA can learn how to discriminate similar legal concepts and how to apply legal knowledge to solve case-analysis questions. Thus, the quality of explanations in the SFT dataset can be important, which means it is improper to train the model with incorrect or uninformative explanations.

To achieve the analysis for each option, we concatenate a question with its four options respectively as four separate inputs for ChatGPT. We also find ChatGPT can generate more coherent responses, if we transfer the concatenated question-answer pair to a single fluent query. We illustrate the transferring process below.

Question: *After one party to the contract has paid the other party compensation for breach of the contract, which of the following options should be taken?*

Option: *It is up to the compensating party to decide whether to continue the performance of the contract.*

We then should transfer this example to the following query: *After one party to the contract has paid the other party compensation for breach of the contract, can the compensating party decide whether to continue the performance of the contract?*

We first utilize regular expressions to transfer the examples in JEC-QA. Among all question-option pairs, 39.9% of them can be processed by our regular expressions, while the rest are removed. We then input the transferred queries into ChatGPT for explanations and answers by chain-of-thought. In this way, we collect 42k examples, denoted as JE-Q2EA.

We further find the conclusions of around 1/4 JE-

Q2EA examples are incorrect, and more than 2/3 of transferred queries are not fluent and unnatural. Therefore, we try to transfer the question-option pair by in-context learning based on ChatGPT. We use the following prompt.

Prompt: *Here is a question from the National Judicial Examination and one of its options. The question texts will begin with "### Question:" and the option texts with "### Option:". Please concatenate them as a single true-or-false question. A few examples are illustrated below.*

Example 1:

Question: [Question 1]

Option: [Option 1]

Output: [A human-annotated result.]

Example 2:

Question: [Question 2]

Option: [Option 2]

Output: [A human-annotated result.]

Example 3:

Question: [Question 3]

Option: [Option 3]

Output: [A human-annotated result.]

Now, please concatenate the following question and option as a single fluent and coherent query. The output should not contain phrases like "options", "from following", "as below".

Question: [Input Question]

Option: [Input Option]

Output:

Almost all of the questions transferred by ChatGPT can keep the same meanings as the original question-option pairs. However, we find there are still many words indicating the queries come from multiple-choice questions, such as the word *the option* in *After one party to the contract has paid the other party compensation for breach of the contract, **the option** "the compensating party can decide whether to continue performance of the contract". Is it correct?"*

Such examples are unnatural and confusing, since there are no contexts about other options. If ChatGPT is given these examples, it might reply "I cannot find other options here, so I cannot answer which option should be chosen." To guarantee the quality of the dataset, we filter out the transferred queries containing words like "option", "from following", etc. In this way, it remains 6k transferred queries. We then input the queries and its answer (True or False) to ChatGPT, and collect reasoning analysis by following the prompt.

Prompt: *Here is a question about law, could you please answer it and give me a detailed analysis?*

Question: [Question]

Directions: *The answer of given question can be "[Yes/No]" or "[Correct/Incorrect]". Please answer it step by step. If the question is about case analysis, please analyze the reasons behind the party's decision to take such action. If the question is about legal concept, please list the legal basis involved in your answer.*

Output:

After examining the replies produced by ChatGPT, we find although the conclusions of most examples are correct, the reasoning analysis might be illogical, which refers to the wrong legal basis. To further improve the quality of the dataset, we collect the analysis of Judicial Examination questions in the past two years, which are written by human experts. We also ask experts to transfer the questions manually. Thanks to their efforts, we can obtain a high-quality dataset of 850 examples.

In Table 6, we show examples of different methods of collecting explanations for JE questions. The answer in the JE-Q2EA example is wrong, let alone its explanation. For the JE-QA2E example, the answer to the question is correct, but the reasoning process is questionable, where a non-existent legal article. The **JE-EXPERT** example is of the best quality, citing the correct law and performing logical reasoning.

A.2 Legal Consultation

To generate the response to a legal consultation question, we ask ChatGPT to act as a lawyer and respond to a client. In the input prompt, we list six requirements that the generated responses should meet: (1) properly citing legal articles, (2) giving well-founded analyses based on the facts of the case and legal articles, (3) responding comprehensively and analyzing the potential possibilities, (4) asking appropriate questions to dig out facts to assist in further answers, (5) using plain language, (6) giving preliminary legal opinions and consulting conclusions.

In the example of Table 5, we retrieve 3 legal articles to facilitate a more faithful generation. Note that two of them are irrelevant to the consultation and ChatGPT is able to ignore them in its response.

B Details of Charge Prediction Tasks

CAIL2018 (Xiao et al., 2018) is large-scale dataset for Judgment Prediction, which is constructed from the criminal documents collected from China Judgments Online². It includes three tasks, to predict the charge, to retrieve related articles, and to predict the terms of penalty. Since most criminal documents mention more than one articles, LLMs cannot give validate predictions zero-shot via perplexity. And previous also reveal that LLMs are not sensitive to numerical texts (Chen et al., 2023). Thus, we do not choose *Article Retrieval* and *Prediction of Terms-of-Penalty* as the evaluation benchmark.

For Charge Prediction, we first filter out the documents which refer to multiple charges. And to construct a challenging task, we select several sets of charges which are easily misidentified to each other. The charges are:

- (1) Forcible seizure (抢夺罪) and Robbery (抢劫罪);
- (2) Theft (盗窃罪) and Criminal Conversion (侵占罪);
- (3) Arson (放火罪) and Arson by Negligence (失火罪);
- (4) Offering Bribes (行贿罪), Acceptance of Bribes (受贿罪) and Embezzlement (贪污罪).

For each charge, we randomly sample 100 documents from CAIL2018. When evaluating, the model should choose a charge from the nine ones as its prediction. We employ zero-shot accuracy as the evaluating metric.

C Examples from Human Evaluation

Here, we analyze some cases from the human evaluation.

In Table 9, we show the responses outputted by the models trained with different JE datasets. In the output by the model trained with JE-Q2EA-42k, there is a misjudgment of the effectiveness of the agreement. The legal act of divorce will only invalidate the marriage relationship, but will not necessarily make the agreement lose the legal validity of the contract. The model trained with JE-QA2E-6k misunderstands Article 1083 of the Civil Code and claims that *if the husband wants to remarry after changing his mind, he can apply to the marriage registration authority to register a new marriage and establish a new marital relationship*. Remarriage requires both parties to jointly

²<https://wenshu.court.gov.cn/>

Related Articles: 《民法典》第一千零四十七条：结婚年龄，男不得早于二十二周岁，女不得早于二十周岁。《民法典》第一千零四十六条：结婚应当男女双方完全自愿，禁止任何一方对另一方加以强迫，禁止任何组织或者个人加以干涉。《民法典》第三百零四条：共有人可以协商确定分割方式。达不成协议...

Client: 我和我对象想结婚，我想问一下，结婚的法定婚龄是几岁啊？

Lawyer: 根据《民法典》第一千零四十七条规定，结婚的法定婚龄为男不得早于22周岁，女不得早于20周岁。如果未到法定结婚年龄而坚持结婚，将被认定为无效婚姻。请问您和您的对象的年龄是否已达到法定婚龄？如果未达到，建议您等待至法定婚龄再考虑结婚的问题。

Related Articles: Article 1,047 of the Civil Code: A man shall attain the age of twenty-two, and a woman the age of twenty to enter into matrimony. Article 1,046 of the Civil Code: A man and a woman shall enter into matrimony freely and voluntarily. Neither party may compel the other to conclude a marriage against the latter's will, and no organization or individual may interfere with the freedom of marriage. Article 304 of the Civil Code: Co-owners may determine the means of partition of their co-owned property by consultation. Where they fail to reach an agreement...

Client: I would like to ask, what is the legal minimum age for marriage?

Lawyer: According to Article 1,047 of the Civil Code, the legal marriageable age is a minimum of 22 years old for males and 20 years old for females. If individuals insist on getting married before reaching the legal marriageable age, the marriage will be deemed void. May I ask if both you and your partner have reached the legal marriageable age? If not, I would suggest waiting until you reach the legal marriageable age before considering marriage.

Table 5: An example of the retrieval augmented legal consultation generated by ChatGPT in Section 2.3.

apply for re-marriage registration, and cannot be applied by one party alone. The biggest advantage of the output by the model trained on JE-EXPERT, compared with the other two editions, is that the validity of the agreement is discussed under different circumstances. The parties did not provide detailed agreement content, but the model infers that there will be agreements on property division, child support, and possibly remarriage in the agreement, showing its reasoning ability. At the same time, the model further infers that remarriage is a voluntary act based on the provisions of the Civil Code and cannot be unilaterally forced to remarry, demonstrating the ability to reason according to the law.

In Table 10, we show an example where legal article retrieval helps alleviate the hallucinations of models. Although the model has been trained on the corpus of legal articles, it cannot recall it accurately when needed. If we ask the model whether cousins can marry without providing relevant legal articles in the input, it will exhibit severe hallucinations: It quotes the wrong number of article and adds non-existent content after a real legal article. When augmented with three relevant articles, the model will give a correct answer based on the correct article, with no hallucinations.

JE-Q2EA

Question: 关于抢劫罪，下列说法错误吗？
驾驶摩托车夺取他人财物时，因被害人不放手而采用强拉硬拽方法劫取财物的，应认定为抢劫罪。
请给出详细的推理过程之后再给出答案。

Answer: 这个说法是错误的。
根据《中华人民共和国刑法》第X条规定，抢劫罪是指以暴力、胁迫方法抢夺公私财物的行为。而强拉硬拽并不属于暴力、胁迫的范畴，因此不符合抢劫罪的构成要件。
此外，即使被害人不放手，也不能因此就认定为采用了暴力、胁迫的手段。因此，驾驶摩托车夺取他人财物时，因被害人不放手而采用强拉硬拽方法劫取财物的行为，应认定为盗窃罪。

Question: Is the following statement about the crime of robbery incorrect? If someone drives a motorcycle and forcibly pulls and drags someone's property because the victim does not let go, it should be recognized as the crime of robbery.

Please provide a detailed reasoning process before giving the answer.

Answer: This statement is incorrect.
According to Article X of the Criminal Law of the People's Republic of China, the crime of robbery refers to the act of forcibly seizing public or private property through violence or coercion. However, forcibly pulling and snatching does not fall within the scope of violence or coercion, and therefore does not meet the criminal elements of the crime of robbery. Furthermore, even if the victim does not let go, it does not automatically imply the use of violence or coercion. Therefore, when someone drives a motorcycle and forcibly pulls and snatches someone's property because the victim does not let go, it should be recognized as the crime of theft.

JE-QA2E

Question: 关于抢劫罪，驾驶摩托车夺取他人财物时，因被害人不放手而采用强拉硬拽方法劫取财物的，应认定为抢劫罪。这个说法是否正确？

Answer: 根据《中华人民共和国刑法》第二百六十四条规定，抢劫罪是指以暴力、胁迫方法抢夺公私财物的行为。而在本题中，驾驶摩托车夺取他人财物时，因被害人不放手而采用强拉硬拽方法劫取财物，显然是采用了暴力、胁迫手段，因此符合抢劫罪的构成要件。此外，根据《中华人民共和国刑法》第六十六条规定，使用暴力、胁迫方法抢夺公私财物，情节严重的，应当判处三年以上、十年以下有期徒刑，并处罚金或者没收财产。因此，对于这里的行为，应当认定为抢劫罪，并依法追究刑事责任。综上所述，这里的正确回答为“正确”，因为该行为符合抢劫罪的构成要件，根据相关法律规定应当认定为抢劫罪。

Question: Is it correct to classify the act of using forceful pulling and dragging methods to seize someone's property while driving a motorcycle as robbery?

Answer: According to Article 264 of the Criminal Law of the People's Republic of China, robbery refers to the act of forcibly taking public or private property through violence or coercion. In this case, when driving a motorcycle to seize someone's property and resorting to forceful pulling and dragging because the victim refuses to let go, it is evident that violent and coercive means are being employed. Therefore, it meets the criminal elements of robbery. Furthermore, according to Article 66 of the Criminal Law of the People's Republic of China, if violence or coercion is used to commit robbery and the circumstances are serious, the offender shall be sentenced to imprisonment for more than three years but less than ten years, in addition to a fine or confiscation of property. Therefore, for this behavior, it should be classified as robbery and the criminal responsibility should be pursued in accordance with the law. In conclusion, the correct answer to this question is "correct" because this behavior meets the criminal elements of robbery and should be classified as such according to relevant legal provisions.

JE-EXPERT

Question: 关于抢劫罪，驾驶摩托车夺取他人财物时，因被害人不放手而采用强拉硬拽方法劫取财物的，应认定为抢劫罪。这个说法是否正确？

Answer: 在刑法理论上，构成抢劫罪的构成要件之一是行为人对被害人当场使用暴力、胁迫或者其他方法。如果犯罪分子没有使用暴力或者胁迫的方法就取得了财物，不能以抢劫罪论处。但是，如果犯罪分子事先只是准备抢夺，但在实施抢夺的过程中遭到反抗或者阻拦，于是当场使用暴力或者以暴力相威胁强取财物，其行为就由盗窃或者抢夺转化为抢劫了，应以抢劫罪定罪处罚。在实体法上，根据《最高人民法院关于审理抢劫、抢夺刑事案件适用法律若干问题的意见》：“驾驶机动车强拉硬拽夺取财物的，因被害人不放手而采取强拉硬拽方法劫取财物的，应当以抢劫罪定罪处罚。”因此，本案定抢劫罪的说法正确。

Question: Is it correct to classify the act of using forceful pulling and dragging methods to seize someone's property while driving a motorcycle as robbery?

Answer: In the theoretical framework of criminal law, one of the criminal elements of the crime of robbery is that the perpetrator employs violence, coercion, or other methods against the victim on the spot. If the criminal obtains property without using violence or coercion, they cannot be charged with robbery. However, if the criminal only intended to commit robbery beforehand but encounters resistance or obstruction during the act of robbery and subsequently employs violence or threatens to use violence to forcibly take the property, their behavior transforms from theft or robbery to robbery, and they should be convicted and punished for the crime of robbery. In terms of substantive law, according to the "Opinions of the Supreme People's Court on Several Issues concerning the Application of Laws in the Trial of Robbery and Extortion Criminal Cases": "When a person drives a vehicle and forcibly snatches property, if they resort to forcibly pulling and dragging to seize the property because the victim does not let go, they should be convicted and punished for the crime of robbery." Therefore, the claim that this case falls under the crime of robbery is correct.

Table 6: Examples of the three different methods for collecting JE explanations in Section 2.2.

请问：民法典中关于“婚姻无效”是如何规定的？	May I ask: How is "void marriage" defined in the Civil Code?	没领证可以生小孩吗？需要什么手续？	Can we have a baby without a marriage certificate? What procedures are required?
大概一年多前，父母帮我们订了婚约，但现在一点感觉都没有了，我能不和他结婚吗？婚约一定得遵守吗？	Approximately over a year ago, my parents helped us arrange an engagement, but now I don't have any feelings at all. Can I not marry him? Do engagements have to be honored?	2022年，甲因胁迫结婚，请求撤销婚姻，应当自胁迫行为终止之日起两年内提出，对吗？	In 2022, Bob was coerced into marriage and now requests the annulment of the marriage. The request should be made within two years from the date of termination of the coercion, right?
邻居两口子为了买房子办了假离婚，还签了协议来着，结果现在女的反悔了，不离婚了。男的不干了，啥也没有了，他们签的这协议有法律效力吗？	The neighbors, a married couple, got a fake divorce to buy a house, and they even signed an agreement. But now the woman has changed her mind and doesn't want to remarry. The man disagrees because he has nothing left. Does the agreement they signed have legal validity?	实在受不了他抽烟喝酒的，还是离了好。也没孩子，其他都谈好了，就差房子了，反正我是没有其他地方住了，这房子我必须要有。他也想要房子，怎么办呢？	I can't stand his smoking and drinking anymore. It's better to get a divorce. We don't have any children, and everything else is settled, except for the house. Anyway, I have nowhere else to live, so I must have this house. He also wants the house, so what should we do?
男方婚内出轨，还和小三生了一个孩子，该怎么办？能让他净身出户吗？	How to deal with it when the husband cheats during the marriage and has a child with the other woman? Can he leave without taking any property after divorce?	终于决定要分居了，虽然还没走到离婚那一步。分居协议书该怎么写呢？	Finally decided to separate, although we haven't reached the divorce stage yet. How should we write a separation agreement?
离婚冷静期是什么意思？	What is divorce cooling-off period?	表兄妹可以结婚吗？	Can cousins marry?
不给孩子的抚养费最严重的后果是什么？	What are the most serious consequences of not providing child support?	我们俩已经分居两年了，算自动离婚吗？	We have been living separately for two years. Does that count as an automatic divorce?
全职母亲离婚时可以分财产吗？	Can a full-time mother get a share of the property in a divorce?	彩礼是夫妻共同财产吗？	Is bride price considered joint property of the husband and wife?
结婚前我爸妈给全款买的房，在城西，涨了一倍还多，现在如果离婚的话，要给她分一半吗？	Before getting married, my parents bought the house with full payment in the western part of the city. It has increased by more than double in value. Now, if we were to divorce, would I have to give her half?	大概5,6年前，我们俩就住一起了，一直也没领证，也没要孩子。现在他老打我，实在受不了了，我能去法院起诉离婚吗？	Approximately 5 or 6 years ago, the two of us started living together, but we never obtained marriage certificate or had children. Now he constantly abuses me, and I can't tolerate it anymore. Can I go to court and file for divorce?
结婚证丢了，我还能去法院起诉他离婚吗？	Can I still sue him for divorce in court if I lost my marriage certificate?	女婿有没有赡养岳父的义务	Does the son-in-law have an obligation to support his father-in-law?
孩子还不满一周，能离婚吗？	The child is not even a week old. Can we get a divorce?	总是被家暴，该怎么保护自己啊？	I am constantly experiencing domestic violence. How can I protect myself?

Table 7: List of the questions for human evaluation. (Part 1)

继父和我妈已经办了离婚手续了，我对他还有赡养义务吗？还要给他抚养费吗？	Stepfather and my mother have already completed the divorce proceedings. Do I still have an obligation to support him? Do I still need to pay alimony to him?	她结婚前也没说自己得了红斑狼疮啊，能离婚吗？	She didn't mention having lupus erythematosus before getting married. Can we get a divorce?
我们俩本来都要结婚了，结果因为彩礼钱黄了，她爸妈又要5万分手费，这合法吗？	We were both planning to get married, but things fell through because of the bride price. Her parents are now demanding 50,000 yuan as a breakup fee. Is this legal?	结婚以后，我爸妈出钱买的房子，只写了我一个人的名字，这算是夫妻共同财产吗？	After getting married, my parents bought a house and only registered my name on it. Is it community property?
我自己都吃不上饭了，还要给养父赡养费吗？	I can't even afford to feed myself. Do I still have to pay support to my foster father?	我是朝鲜族人，今年虚岁20了，可以领结婚证吗？	I am of Korean ethnicity, and I am 20 years old by traditional reckoning. Can I apply for a marriage certificate?
复婚和再婚有什么区别吗？	What is the difference between remarriage and second marriage?	同姓的人可以结婚吗？	Can people with the same surname get married?
领养小孩需要啥条件？	What are the requirements for adopting a child?	军人多少岁可以结婚？	At what age can military personnel get married?
30岁领证算晚婚吗？	Is getting married at 30 considered a late marriage?	离婚多久能复婚？	How long after a divorce can you remarry?
刚离婚就去登记算重婚吗？	it considered bigamy if you register another marriage immediately after getting divorced?	离异再婚要改户口本吗？	Do you need to update the household registration for remarriage after divorce?
再婚的时候怎么保护婚前个人财产	How can I protect my personal property acquired before marriage when remarrying?	没有户口本能领结婚证吗？	can you get a marriage certificate without a household registration?
结婚证丢了怎么办？	What should I do if I lost my marriage certificate and need to renew it?		

Table 8: List of the questions for human evaluation. (Part 2)

Question: 邻居两口子为了买房子办了假离婚，还签了协议来着，结果现在女的反悔了，不离婚了。男的不干了，啥也没有了，他们签的这协议有法律效力吗？

Articles: 《民法典》第一千零八十条：完成离婚登记，或者离婚判决书、调解书生效，即解除婚姻关系。

《民法典》第一千零八十三条：离婚后，男女双方自愿恢复婚姻关系的，应当到婚姻登记机关重新进行结婚登记。

《民法典》第一千零九十八条：收养人应当同时具备下列条件：（一）无子女或者只有一名子女；（二）有抚养、教育和保护被收养人的能力；（三）未患有在医学上认为不应当收养子女的疾病；（四）无不利于被收养人健康成长的违法犯罪记录；（五）年满三十周岁。

JE-EXPERT: 您好，根据《民法典》的规定，完成离婚登记或者离婚判决书、调解书生效，即解除婚姻关系。因此，如果邻居两口子已经办理了离婚登记手续，那么他们的婚姻关系已经解除。如果他们签署了协议，该协议是否具有法律效力需要看具体情况。首先，如果该协议是关于离婚后的财产分割、赡养子女等方面的约定，那么该协议是有效的。因为夫妻离婚后，财产分割、赡养子女等问题需要通过协商达成一致意见，并签订书面协议。如果没有违反法律规定，该协议是具有法律效力的。其次，如果该协议是关于复婚方面的约定，那么该协议不一定有效。因为复婚是夫妻自愿的行为，如果女方不同意复婚，男方单方面无法强制复婚。最后，建议您了解一下具体的情况，比如该协议的具体内容、签订时间以及是否符合相关法律法规的规定等。如果您需要进一步的帮助，可以向当地的律师咨询。

JE-Q2EA-42k: 您好，根据《民法典》第一千零八十条规定，完成离婚登记或者离婚判决书、调解书生效，即解除婚姻关系。因此，如果两人已经办理了离婚手续并签署了协议，那么这个协议就无效了。但是，如果这个协议是关于房屋买卖的相关事宜，可以根据合同法的规定来解决。请问这个协议的具体内容是什么？是否涉及到房屋买卖等相关事宜？

JE-QA2E-6k: 您好，根据《民法典》第一千零八十条规定，完成离婚登记，或者离婚判决书、调解书生效，即解除婚姻关系。因此，如果两人已经办理了离婚手续并签署了离婚协议，那么这份协议就生效了，具有法律效力。但是如果男方反悔想要复婚，那么他可以向婚姻登记机关申请重新进行结婚登记，重新建立婚姻关系。请问你们之间是否已经完成离婚登记呢？如果没有完成离婚登记，那么这份协议就没有生效，也就不存在法律效力。

Question: The neighbors, a married couple, got a fake divorce to buy a house, and they even signed an agreement. But now the woman has changed her mind and doesn't want to remarry. The man disagrees because he has nothing left. Does the agreement they signed have legal validity?

Articles: Article 1,080 of the Civil Code: A marital relationship is dissolved on completion of registration of divorce or when a judgment granting divorce or a mediation agreement on divorce takes effect.

Article 1,083 of the Civil Code: Where, after divorce, both the man and the woman voluntarily resume their marital relationship, they shall once again file for registration of marriage at a marriage registry.

Article 1,098 of the Civil Code: A prospective adopter shall meet all of the following criteria: 1. Having no child or only one child; 2. Being capable of raising, educating and protecting the adoptee; 3. Being free of any disease which renders the prospective adopter medically unfit to be an adopter; 4. Having no criminal record unfavorable to the healthy growth of the adoptee; and 5. Having attained the age of thirty.

JE-EXPERT: Hello, according to the provisions of the Civil Code, completing the registration of divorce or the effective divorce judgment or mediation agreement will dissolve the marriage relationship. Therefore, if your neighbors have already completed the divorce registration procedures, their marriage relationship has been dissolved. Whether their agreement has legal effect depends on the specific circumstances. Firstly, if the agreement pertains to the division of property and child support after divorce, then the agreement is valid. After a couple divorces, issues such as property division and child support need to be agreed upon through negotiation and documented in a written agreement. If the agreement does not violate any legal provisions, it has legal effect. Secondly, if the agreement pertains to remarriage, then the agreement may not be valid. Remarriage is a voluntary act between spouses, and if one party does not agree to remarry, the other party cannot force them to do so unilaterally. Lastly, it is advisable to gather specific information about the situation, such as the exact content of the agreement, the date of signing, and whether it complies with relevant laws and regulations. If you need further assistance, you can consult a local lawyer.

JE-Q2EA-42k: Hello, according to Article 1080 of the Civil Code, the divorce registration, divorce judgment, or mediation agreement becomes effective and terminates the marital relationship. Therefore, if both parties have completed the divorce procedures and signed an agreement, that agreement is considered invalid. However, if the agreement pertains to matters related to property transactions, such as the sale of a house, it can be resolved in accordance with the provisions of the Contract Law. Could you please provide the specific content of this agreement? Does it involve matters related to the sale of a house or other related issues?

JE-QA2E-6k: Hello, according to Article 1080 of the Civil Code, completing the divorce registration or having the divorce judgment or mediation agreement take effect will terminate the marital relationship. Therefore, if both parties have already gone through the divorce procedures and signed the divorce agreement, then this agreement is valid and has legal effect. However, if the husband wants to remarry after changing his mind, he can apply to the marriage registration authority to register a new marriage and establish a new marital relationship. May I ask if you have already completed the divorce registration between you? If the divorce registration has not been completed, then this agreement is not effective and does not have legal force.

Table 9: Examples of the responses outputted by the models trained with different JE datasets in Section 5.1.

<p>Question: 表兄妹可以结婚吗?</p> <p>Articles: 《民法典》第一千零四十八条: 直系血亲或者三代以内的旁系血亲禁止结婚。 《民法典》第一千零五十一条: 有下列情形之一的, 婚姻无效: (一) 重婚; (二) 有禁止结婚的亲属关系; (三) 未到法定婚龄。 《民法典》第一千零四十九条: 要求结婚的男女双方应当亲自到婚姻登记机关申请结婚登记。符合本法规定的, 予以登记, 发给结婚证。完成结婚登记, 即确立婚姻关系。未办理结婚登记的, 应当补办登记。</p>	<p>Question: Can cousins marry?</p> <p>Articles: Article 1048 of the Civil Code: Persons who are lineal relatives by blood, or collateral relatives by blood within three generations are prohibited from marrying each other. Article 1051 of the Civil Code: A marriage is null and void under any of the following circumstances: 1. Bigamy; 2. The parties to the marriage are within the prohibited degrees of relationship; or 3. Either party to the marriage has not yet attained the legal marriage age. Article 1049 of the Civil Code: Both the man and the woman intending to enter into matrimony shall apply for registration of the marriage in person with a marriage registry. If the marriage proposed is found to conform to the provisions of this Code, the marriage shall be registered, and a marriage certificate shall be issued. A marital relationship shall form on the completion of the marriage registration. A couple who has failed to file a marriage registration shall complete the registration.</p>
<p>w/ Hallucination: 《婚姻法》第二十一条第一款规定: “有下列情形之一的, 为无效婚姻: (一) 重婚的; (二) 有禁止结婚的亲属关系的; (三) 未到法定婚龄的; (四) 被法律认定为无民事行为能力人或者限制民事行为能力人的; (五) 在婚姻登记过程中弄虚作假, 骗取婚姻登记的; (六) 其他违法婚姻的。”据此, 表兄妹是三代以内旁系血亲, 属于《民法典》禁止结婚的亲属关系, 因此, 他们不能结婚。</p>	<p>w/ Hallucination: Article 21, paragraph 1 of the Marriage Law states: "A marriage shall be void under any of the following circumstances: (1) bigamy; (2) marriage between relatives within the prohibited degrees; (3) marriage before reaching the legal marriage age; (4) marriage involving a person who has been recognized by law as lacking capacity for civil conduct or whose capacity for civil conduct is restricted; (5) obtaining marriage registration through fraudulent means during the marriage registration process; (6) other illegal marriages." Based on this, first cousins are blood relatives within the third degree of collateral consanguinity, which falls under the prohibited degrees of marriage according to the Civil Code. Therefore, they cannot get married.</p>
<p>w/o Hallucination: 表兄妹是三代以内的旁系血亲, 因此是不能在民法典范围内确定婚姻关系的。</p>	<p>w/o Hallucination: Cousins are collateral blood relatives within three generations, therefore their marriage relationship cannot be established within the scope of the Civil Code.</p>

Table 10: An example of (1) the model output with hallucinations when not provided with legal articles and (2) the model output without hallucinations when augmented with retrieved legal articles in Section 5.2.