

DISC-LawLLM: Fine-tuning Large Language Models for Intelligent Legal Services

Shengbin Yue^{1,2}, Wei Chen³, Siyuan Wang⁴, Bingxuan Li⁴, Chenchen Shen⁴, Shujun Liu⁴, Yuxuan Zhou⁴, Yao Xiao⁷, Song Yun⁶, Xuanjing Huang⁵, Zhongyu Wei^{1,4*}

¹Research Institute of Intelligent Complex Systems, Fudan University, China

²Shanghai Center For Mathematical Sciences, Fudan University, China

³School of Software Engineering, Huazhong University of Science and Technology, China

⁴School of Data Science, Fudan University, China

⁵School of Computer Science, Fudan University, China

⁶Rule of Law Institute, Northwest University of Political and Law, China

⁷New York University Shanghai, China

{sbyue23,bxli16,ccshen22,yxzhou23}@m.fudan.edu.cn, 1171991@s.hlju.edu.cn, yx2436@nyu.edu, {chenwei18,wangsy18,shujunliu20,xjhuang,zywei}@fudan.edu.cn

Abstract

We propose DISC-LawLLM, an intelligent legal system utilizing large language models (LLMs) to provide a wide range of legal services. We adopt legal syllogism prompting strategies to construct supervised fine-tuning datasets in the Chinese Judicial domain and fine-tune LLMs with legal reasoning capability. We augment LLMs with a retrieval module to enhance models' ability to access and utilize external legal knowledge. A comprehensive legal benchmark, DISC-Law-Eval, is presented to evaluate intelligent legal systems from both objective and subjective dimensions. Quantitative and qualitative results on DISC-Law-Eval demonstrate the effectiveness of our system in serving various users across diverse legal scenarios. The detailed resources are available at <https://github.com/FudanDISC/DISC-LawLLM>.

1 Introduction

With the rise of legal artificial intelligence (LegalAI) (Gardner, 1987; Zhong et al., 2020a), the legal domain is undergoing significant transformation. Through automating legal tasks including legal information extraction (Bommarito et al., 2018), interactive argument pair extraction (Ji et al., 2018, 2019; Yuan et al., 2021), case retrieval (Ma et al., 2021), judgment prediction (Song and Wei, 2021; Ye et al., 2018; Yang et al., 2019), and legal question answering (Kien et al., 2020), intelligent legal systems benefit various groups of people. It boosts the efficiency of legal professionals by reducing the heavy burden of paperwork, and simpli-

fies access to legal services and remote legal advice for general populations. Besides, it offers invaluable assistance to students in their legal knowledge pursuits and examinations.

Prior work on LegalAI mainly focuses on specific tasks by constructing datasets and presenting machine learning algorithms. Although they demonstrate some impressive results, this approach confines the broader application of intelligent legal systems. Recent advancements in large language models (LLMs) (OpenAI, 2022, 2023; Muenighoff et al., 2023; Meta, 2023) have shown remarkable instruction-following ability across varied domains, offering tremendous potential for the legal sector. Some initial progress has been made (Huang et al., 2023; Cui et al., 2023a) by fine-tuning general LLMs to utilize legal knowledge for simple question answering, however, legal services are far more complicated and beyond dialogue.

As shown in Figure 1, intelligent legal systems have broad application scenarios, serving different groups of users, including professional legal practitioners, everyday individuals seeking legal advice, and law students pursuing academic achievement. For legal practitioners, the system should provide advanced legal tools for statute retrieval, case analysis, and document summarization. For the general public, the system should be able to offer legal consultation for statute interpretation and dispute resolution. For law students, the system serves as a tutor, helping to consolidate legal knowledge and providing solutions to exam questions.

An example for legal consultation is illustrated in Figure 1, where the intelligent legal system can leverage law knowledge to mine facts from the in-

*Corresponding Author

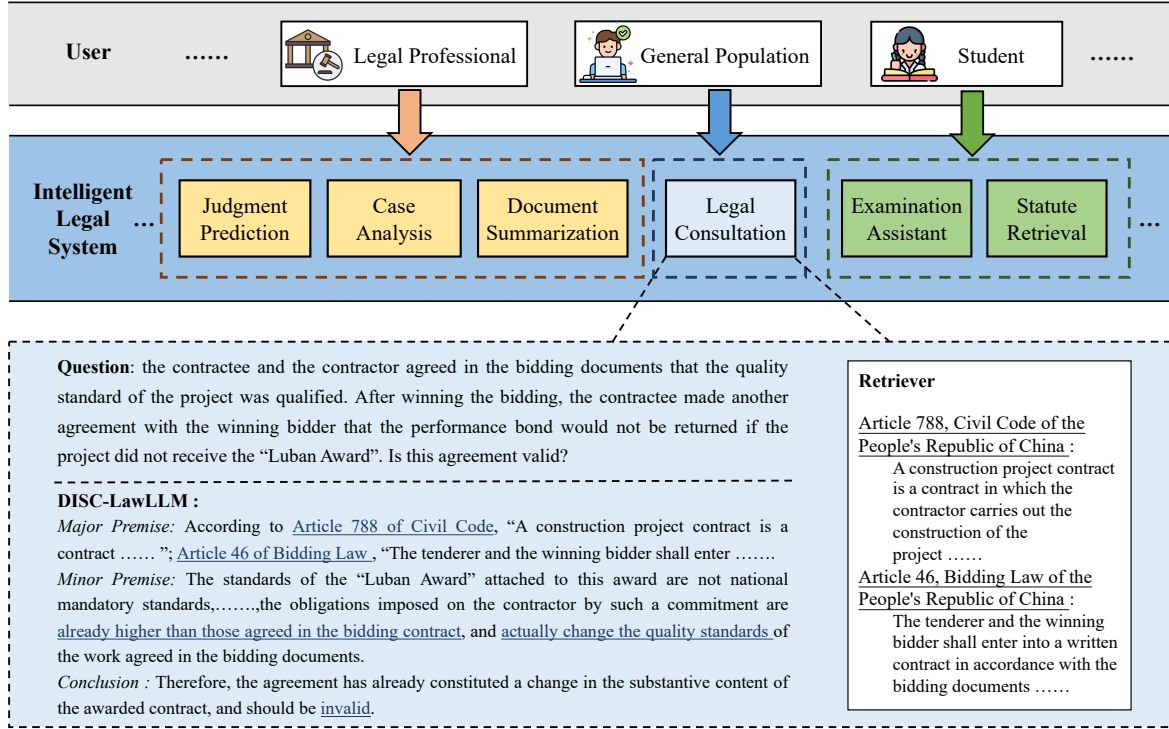


Figure 1: Overview of DISC-LawLLM serving different users in different legal scenarios. For an example of legal consultation, DISC-LawLLM utilizes the retrieved statutes to perform legal reasoning (i.e., the statute is the major premise, the fact in the question is the minor premise, and the final inference is the conclusion).

quiry, and deduce the conclusion to provide legal services. This highlights two primary challenges for building the system. (1) High demand of reasoning ability for legal issues. The legal assistance process is specialized and requires intricate legal reasoning. For example, all legal responses in jurisprudence should follow the structure of legal syllogism (Posner, 1990; Jiang and Yang, 2023), which involves a major premise representing the legal proposition, a minor premise symbolizing the factual proposition, and a conclusion representing the judgment. (2) In need of retrieval and inference ability of external legal knowledge. It requires a deep understanding and precise reference of legal knowledge for delivering legal interpretation and alleviating the hallucination. Besides, the repository of legal knowledge is constantly evolving with newly emerged and amended regulations.

To this end, we present DISC-LawLLM, our legal large language model tailored for building intelligent legal systems with legal reasoning and knowledge retrieval capability. We begin by adopting the legal syllogism prompting strategy to construct supervised fine-tuning datasets in the Chinese Judicial domain, named DISC-Law-SFT. These datasets are then employed to train DISC-LawLLM

with legal reasoning on top of a general domain Chinese LLM with 13B parameters¹. Besides, we also introduce a retrieval module to source up-to-date and precise legal evidence, enhancing the retrieval augmented DISC-LawLLM’s ability to generate more reliable responses.

Finally, we design a legal benchmark, DISC-Law-Eval, to provide a comprehensive assessment of intelligent legal systems from both objective and subjective dimensions. For the objective perspective, our benchmark incorporates multiple-choice with both single and multiple answers sourced from law-related standardized examinations. These questions are categorized into three difficulty levels: *Easy*, *Normal* and *Hard*, which allows for a deeper insight into the model’s grasp of legal knowledge and reasoning capabilities. For the subjective perspective, we curate a select compilation of high-quality legal Q&A cases, and utilize GPT-3.5 as an arbitrator to assess the model’s metrics in terms of *Accuracy*, *Completeness* and *Clarity*. For each evaluation question, we provide a ground truth to the arbitrator model to reduce potential biases during

¹In this version, we use Baichuan (Baichuan-inc, 2023) as the base model. Note that our strategy can be applied to all decoder-only foundation models.

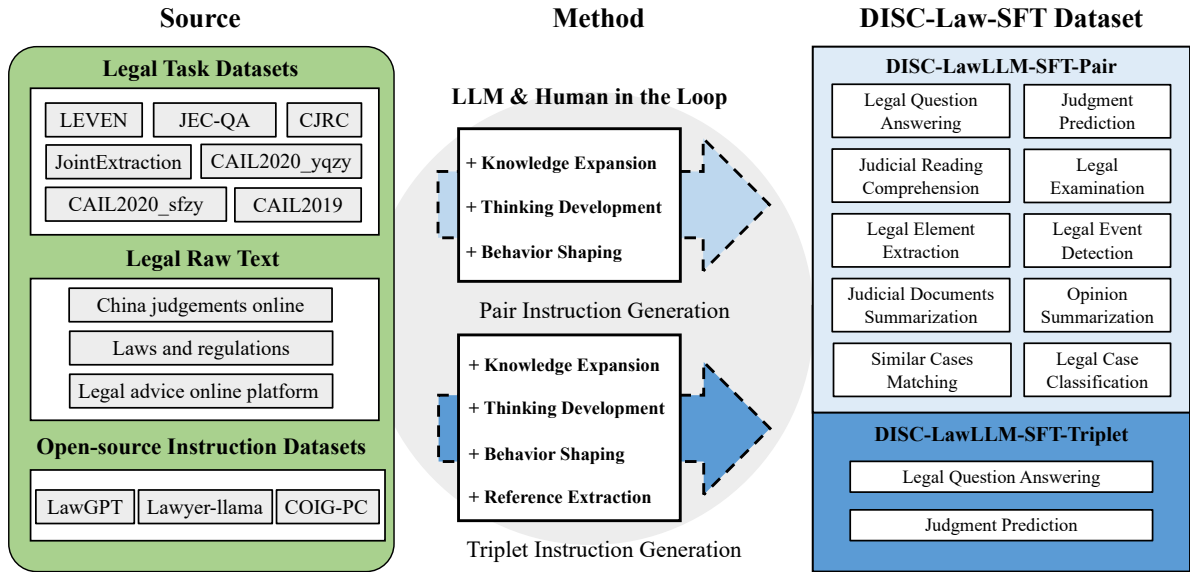


Figure 2: Construction of DISC-Law-SFT Datasets

the assessment phase.

Experimental Results reveal that DISC-LawLLM significantly outperforms existing legal large language models. Even compared to GPT-3.5-turbo (OpenAI, 2022) with 165B parameters, DISC-LawLLM excels in a majority of tested subjects of objective evaluation. Our DISC-LawLLM equipped with more extensive Chinese legal knowledge and legal reasoning, consistently generates more reliable responses.

2 Related Work

Large Language Models (LLMs) have achieved astounding performance on different conventional linguistic tasks, demonstrating powerful generality. However, these generic LLMs have proven to be unsuitable for some domain-specific tasks, such as law. This has greatly stimulated researchers' enthusiasm to explore LLMs in the legal domain. Currently, some initial progress has been made in legal LLMs. Specifically, the LaWGPT (Song, 2023) series was built on Chinese-LLaMA-7B (Cui et al., 2023b), ChatGLM (Du et al., 2022), and Chinese-llama-plus-7B (Cui et al., 2023b) by training with integrated datasets from the Chinese legal domain and a large-scale Chinese legal corpus enriched with domain-specific terminologies. Lawyer LLaMa (Huang et al., 2023) conducted continuous pre-training on Chinese-LLaMA-13B (Cui et al., 2023b) and constructed a large number of instruction finetuning datasets to further enhance its ability to provide legal advice. Lex-

iLaw, based on ChatGLM-6B (Du et al., 2022), was trained with three different methods including LoRA, P-tuning, and finetuning. Additionally, LawGPT_zh (Song, 2023) used self-instruct methods to construct a Q&A dataset and used LoRA to fine-tune ChatGLM-6B. Chatlaw (Cui et al., 2023a) was trained based on Ziya-LLaMA-13B-v1 (IDEA-CCNL, 2021) and Anima-33B separately. Previous work has focused on dialogue competence, one of the intelligent justice tasks. Different from them, we propose an intelligent legal system to provide a wide range of legal services.

3 DISC-Law-SFT Datasets

To train DISC-LawLLM, we construct a high-quality supervised fine-tuning dataset, DISC-Law-SFT with two subsets, namely DISC-Law-SFT-Pair and DISC-Law-SFT-Triplet. The former part aims to introduce the legal reasoning ability to the LLM, while the later part help to improve the model's ability of utilizing external knowledge. The workflow of constructing DISC-Law-SFT is shown in Figure. 2.

Legal intelligent applications in different scenarios usually require combinations of multiple fundamental capabilities of legal text understanding and generating. To this end, we construct instruction samples converging a range of legal tasks, including legal information extraction, judgment prediction, document summarization, and legal question answering, ensuring coverage of diverse scenarios. General LLMs and human labeller are involved to

Dataset	Task	Size	Scenario
DISC-Law-SFT-Pair	Legal Element Extraction	32K	Legal Professional Tools
	Legal Event Detection	27K	
	Legal Case Classification	20K	
	Judgement Prediction	11K	
	Similar Cases Matching	8K	
	Documents Summarization	9K	Legal Consultation
	Public Opinion Summarization	6K	
	Legal Question Answering	93K	
	Document Reading Comprehension	38K	
DISC-Law-SFT-Triplet	Judicial Examination	12K	Examination Assistant
	Judgement Prediction	16K	Legal Professional Tools
	Legal Question Answering	23K	Legal Consultation
General	Alpaca-GPT4	48K	General
	Firefly	60K	
Total		403K	

Table 1: Statistics of DISC-Law-SFT Dataset.

re-construct original samples to generate instructions in two forms of the pair (<input, output>) and the triplet (<input, output, reference>).

3.1 Data Sources

We obtain original samples from three sources, namely, public NLP legal task datasets, legal raw text and open-source instruction datasets.

1) Public NLP Legal Task Datasets. Public datasets covers a range of legal NLP tasks and provide human annotations which can be utilized to generate high quality instructions. We collect public datasets of specific legal tasks related to Chinese justice, including Legal Information Extraction (LEVEN (Yao et al., 2022) and JointExtraction (Chen et al., 2020)), Legal Text Summarization (CAIL2020-sfzy (CAIL, 2020) and CAIL2022-yqzy (CAIL, 2022)), Legal Question Answering (JEC-QA (Zhong et al., 2020b) and CJRC (Duan et al., 2019)), and Judgement Prediction (CAIL2018 (Xiao et al., 2018)).

2) Legal Raw Text. In order to include more scenarios of legal services, we explore to generate instructions from legal raw text. We crawl up an expansive collection of real-world legal text to construct instruction data. This includes consultation data from judicial advisory websites, Chinese laws and regulations, typical cases, judicial verdicts and law-related examinations.

3) Open-source Instruction Datasets. In addition, we also borrow some samples from re-

cently opened instruction datasets. We collect open-source instruction data, including Lawyer-LLaMa (Huang et al., 2023), LawGPT-zh (Liu et al., 2023) and COIG-PC (Zhang et al., 2023).

3.2 Pair Instruction Generation

To construct instructions for supervised fine-tuning DISC-LawLLM, we first use rule-based methods to clean the data and transform it into “input-output” pairs. However, these pairs are too rigid and noisy in linguistic patterns and the expression styles can differ across sources. Therefore, we reconstruct the instruction pairs using the following three methods with the assistance of general large language model.

Behavior Shaping. In the syllogism of legal judgment, the major premise is the applicable law, while the minor premise is pertinent facts, and the conclusion is the final judgment. This constitutes a foundational legal reasoning process for judges. Every case can culminate in a conclusion articulated through a syllogism, as outlined below:

- Major premise: laws
- Minor premise: pertinent facts
- Conclusion: judgment

Inspired by legal syllogism prompting (Jiang and Yang, 2023) and self-construct (Wang et al., 2022), we utilize LLMs to refine output responses for consistency with legal syllogism. We design prompts for GPT-3.5-turbo, to ensure that each conclusion should be drawn from laws and pertinent facts, and responses should be in Chinese.

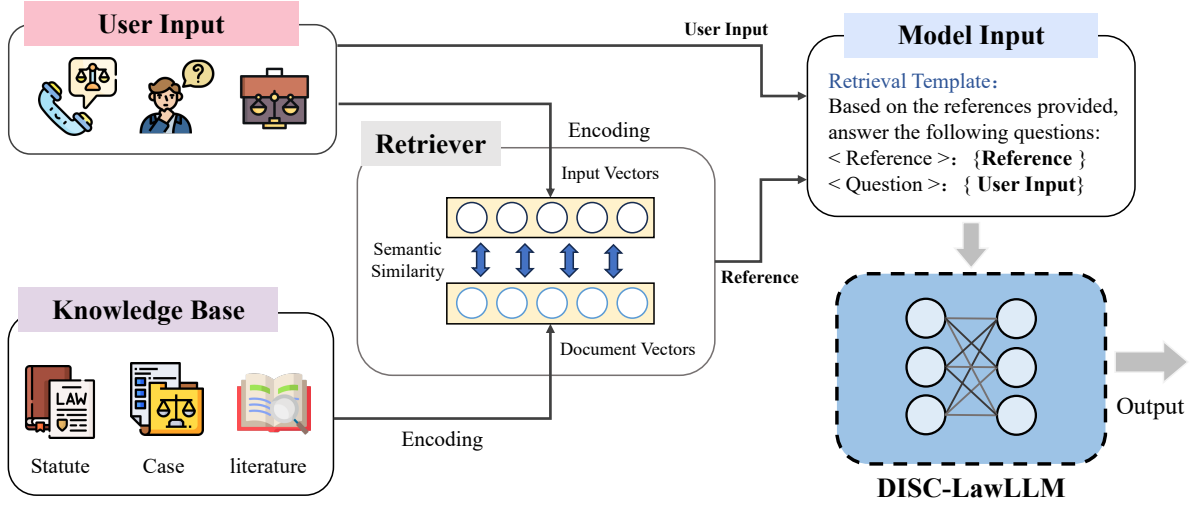


Figure 3: Overview of Retrieval Augmented DISC-LawLLM. Specifically, the reference related to user input in the knowledge base is first retrieved, and then the reference and user input are fed into the DISC-LawLLM with retrieval behavior.

Knowledge Expansion. For multiple-choice questions where behavior shaping is not applicable for selecting an option, we directly expand output responses with legal knowledge to provide more reasoning details. These questions come from various Chinese law-related exams and knowledge competitions, involving knowledge of criminal law, constitutional law, and civil law. While many of them only offer answer options, we use LLMs to expand the involved legal knowledge given the correct answer and reconstruct instruction pairs.

Thinking Development. Chain of Thought (CoT) has been proven effective in enhancing the reasoning ability of models. To further endow legal reasoning into the model, we devise law-specific chains of thought, termed LCoT, to enforce the model conduct legal syllogism to derive the answer. LCoT incorporates prompts that transform input X into X_l as follows:

In the legal syllogism, the major premise is articles of law, the minor premise is the facts of the case, and the conclusion is the judgment of case.

Case: X

Let us use legal syllogism to think and output the judgment:

3.3 Triplet Instruction Generation

To generate supervised instruction triplets <input, output, reference> for retrieval augmented DISC-LawLLM, we create a subset called DISC-Law-SFT-Triplet. For each entry, we utilize the three

strategies outlined in Sec. 3.2 to process the original data and obtain the input and output. Subsequently, we design heuristic rules to extract reference information from this raw data.

3.4 Dataset Overview

Our DISC-Law-SFT dataset consists of more than 10 tasks, such as Legal Element Extraction, Case Matching, Judgment Prediction, Document Summarization, and Question Answering, covering a diverse range of legal scenarios. Additionally, we incorporate general instruction data to enrich the diversity of our training set, mitigating the risk that foundational capability diminishes during the SFT training phase in the legal domain. Specifically, we sourced over 100k samples from alpaca_gpt4_data_zh (Peng et al., 2023) and Firefly (Yang, 2023). Detailed statistics of our datasets are provided in Table 1.

4 DISC-LawLLM

To build an intelligent legal system with legal reasoning and retrieval ability, we form our DISC-LawLLM using two steps, Supervised Fine-Tuning (SFT) and Retrieval Augmentation.

4.1 Supervised Fine-Tuning

We first develop our DISC-LawLLM on top of the Baichuan-13B-Base model (Baichuan-inc, 2023), which is an open-source LLM with over 13.2 billion parameters that was trained on 1.4 trillion tokens corpus, exhibiting ideal performance in both

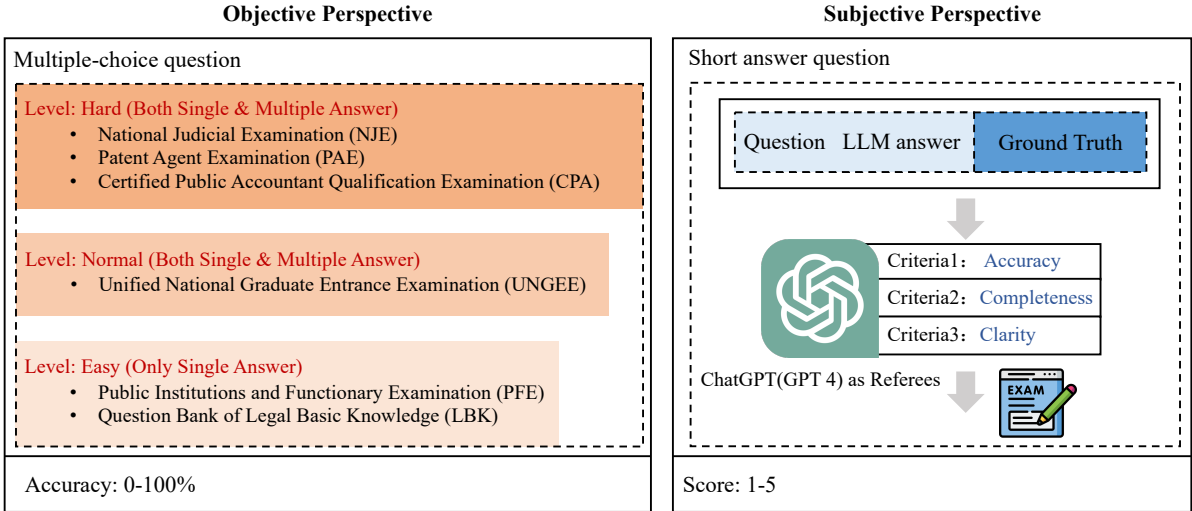


Figure 4: Overview of DISC-Law-Eval Benchmark, assessing systems from both objective and subjective perspectives.

English and Chinese. Specifically, we perform supervised fine-tuning using our DISC-Law-SFT dataset. This refined SFT data enabled the model to be equipped with legal reasoning and judicial behavioral patterns.

The hyperparameters setting of this training process are as follows: global batch size of 256, learning rate of $5e-5$, 2 epochs training stage, maximum source length of 2048 tokens, maximum target length of 1024 tokens. The training process was carried out on 8*A800 GPUs and the training cost is further reduced with the help of deepspeed (Rasley et al., 2020).

4.2 Retrieval Augmentation

In many legal scenarios, such as legal consultation and judgment prediction, users expect model’s responses to be strongly supported by legal precedents and statutes. While we fine-tune LLM with high-quality instruction data, it might produce inaccurate responses due to hallucinations or outdated knowledge. To address this, we augment the DISC-LawLLM with a retrieval module based on an open-source retrieval framework².

We begin by building a knowledge base with over 50 categories of Chinese laws, including the Constitution, Criminal Law, Administrative Procedure Law, Copyright Law, Patent Law. We encode these laws as vectors and save them locally. Given a user input, our retriever then returns Top-K most relevant documents from the knowledge base

by calculating their similarity to the input. These candidate documents, along with the user input, are formulated using our designed template and then fed into the DISC-LawLLM. By querying the knowledge base for references, the model can better understand the major premise, leading to more accurate and reliable answers.

To adapt to retrieval scenarios, we specifically employ DISC-Law-SFT-Triplet, as mentioned in Section 3.3, as our SFT dataset for training. This enables the model to infer reliable results using retrieved references. In addition, our knowledge base is designed for dynamic updates, ensuring the availability of up-to-date laws. Therefore, our thinking-developed DISC-LawLLM can deduce the correct answer based on the new knowledge retrieved.

5 DISC-Law-Eval Benchmark

There is no established benchmark to provide a comprehensive assessment of intelligent legal systems. Inspired by the composition of the bar exam, as shown in Figure 4, we develop a fair evaluation framework, DISC-Law-Eval Benchmark, assessing systems from both the objective perspective and subjective perspective.

5.1 Objective Evaluation

To objectively and quantitatively assess the legal knowledge and reasoning capabilities of intelligent legal systems, we design an objective evaluation dataset. It consists of multiple-choice questions, and each may have one or multiple correct answers.

²<https://github.com/chatchat-space/Langchain-Chatchat>

Model	Size	Hard						Normal		Easy		Average
		NJE		PAE		CPA		UNGEE		PFE	LBK	
		S	M	S	M	S	M	S	M	S	S	
ChatGLM	6B	31.66	1.08	27.97	2.90	37.06	13.33	39.69	20.69	37.65	42.91	24.66
Baichuan-Chat	13B	31.47	10.15	29.66	8.70	35.53	19.17	50.0	27.59	53.12	53.45	30.78
Chinese- <i>alpaca</i> 2	13B	25.7	10.15	30.51	11.59	32.99	19.17	40.94	21.84	44.12	43.27	26.73
GPT-3.5-turbo	175B	36.5	10.58	37.29	17.03	42.13	21.67	51.25	28.74	53.53	54.18	34.10
LexiLaw	6B	20.11	7.56	23.73	10.14	24.87	19.17	31.56	16.09	31.76	40.36	21.50
LawGPT	7B	22.91	6.26	31.36	7.61	25.38	16.67	30.31	13.79	34.71	29.09	20.60
Lawyer LLaMA	13B	35.75	5.62	32.20	6.52	29.95	13.33	32.50	14.94	39.41	39.64	25.05
ChatLaw	13B	27.56	7.99	31.36	9.42	35.53	11.67	35.62	17.24	42.35	41.09	25.20
DISC-LawLLM	13B	42.09	19.87	40.68	18.48	39.59	19.17	50.94	25.29	57.06	54.91	37.10

Table 2: Results compared with general and legal LLMs on Objective Evaluation. Bold represents the best result and underlining represents the second best result. S and M are shorthand of single-answer and multiple answers, respectively.

Subject	Level	S	M	Total
CPA	Hard	197	120	317
NJE		537	463	1000
PAE		118	276	394
UNGEE	Normal	320	87	407
LBK	Easy	275	-	275
PFE		170	-	170

Table 3: Details of Objective Question Dataset, where S and M are shorthand of single-answer and multiple answers, respectively.

It can provide a more challenging and reliable measure of whether the model can use its knowledge to reason toward correct answers. We calculate the accuracy to indicate the performance.

We collect multi-choice questions from a range of Chinese legal standardized examinations and knowledge contests, including National Judicial Examination (NJE), Patent Agent Examination (PAE), Certified Public Accountant Qualification Examination (CPA), Unified National Graduate Entrance Examination (UNGEE), Public Institutions and Functionary Examination (PFE) and Question Bank of Legal Basic Knowledge (LBK). According to content complexity and deduction difficulty, we categorize these questions into three levels: *Hard*, *Normal* and *Easy*. Considering that many legal LLMs use JEC-QA(Zhong et al., 2020b) (2007-2017 National Judicial Examination) for their training datasets, our NJE contains a manual collection of test questions during 2018-2022, ensuring a fair evaluation. Table 3 shows the details of the objective question dataset.

We conduct the objective evaluation in a few-shot setting (4-shot for single-answer questions and 5-shot for multi-answer questions). We use a regular matching method to extract answers from the LLM output, and subsequently compare them to the ground truth to calculate accuracy.

5.2 Subjective Evaluation

We further conduct a subjective evaluation to explicitly demonstrate the model’s command over legal knowledge and reasoning ability. We adopt a question-answering paradigm for this assessment, simulating the process of subjective examination questions. We manually construct a high-quality test set from legal consultations, online postings, justice-related publications, and legal documents, comprising 300 examples. These examples cover scenarios including legal tools, legal consultations, and judgment prediction.

To evaluate this subjective response, we evaluate the model’s output by eliciting a referee model. Strong LLM judges like GPT-3.5, GPT-4 align well with controlled and crowdsourced human preferences (Zheng et al., 2023). In our evaluation, GPT-3.5 serves as a referee and performs the evaluation by providing a rating score from 1 to 5 for each of the following three criteria: *accuracy*, *completeness* and *clarity*.

- Accuracy: The content and semantics of the pending scored answer should be consistent with reference answer.
- Completeness: Compared to the reference answer, the pending scored answer does not miss

Model	Size	ACC	CPL	CLR	Average
ChatGLM	6B	2.64	2.75	3.23	2.87
Baichuan-Chat	13B	3.22	3.34	3.18	3.25
Chinese-Alpaca2	13B	3.13	3.23	3.17	3.17
LexiLaw	6B	3.06	2.62	3.00	2.90
LaWGPT	7B	3.02	2.58	2.96	2.86
Lawyer-LLaMa	13B	3.13	2.83	3.35	3.10
ChatLaw	13B	3.31	2.90	3.35	3.19
DISC-LawLLM	13B	3.46	3.12	3.59	3.39

Table 4: Results compared with general and legal LLMs on Subjective Evaluation, where ACC, CPL and CLR are the shorthand of Accuracy, Completeness and Clarity, respectively.

any details in the reference answer. Do not let the length of the pending scored answer influence your judgment.

- Clarity: Compared to the reference answer, the juridical logic analysis of the pending scored answer is rigorous and clear, and the sentences are well-organized.

To reduce the self-bias of the referee model, we provide the ground truth to the referee as well, enabling them to score according to the ground truth. We repeat the scoring for each question and finally get the average score on different dimensions.

6 Experiments

To demonstrate the excellence of our model, we compare DISC-LawLLM (without retrieval augmentation) with instruction-aligned general LLMs and exiting legal LLMs on the DISC-Law-Eval benchmark. The instruction-aligned general LLMs includes: 1) GPT-3.5-Turbo (OpenAI, 2022); 2) Chatglm-6B (Du et al., 2022); 3) Baichuan-13B-Chat (Baichuan-inc, 2023); 5) Chinese-Alpaca2-13B (ymcui, 2023). The legal LLMs include: 1) LaWGPT (Song, 2023); 2) Lawyer-LLaMa (Huang et al., 2023); 3) ChatLaw (Cui et al., 2023a); 4) LexiLaw (Li, 2023).

6.1 Results in Objective Evaluation

Table 2 shows the objective evaluation performance. We can see that DISC-LawLLM surpasses nearly all competing LLMs across all subjects with different difficulty levels. Even compared to GPT-3.5-Turbo (OpenAI, 2022) with 175B parameters, DISC-LawLLM consistently demonstrates superior performance on most subjects, improving accuracy by an average of 7%. This illustrates the effectiveness of our DISC-LawLLM in inferring correct

answers from questions across a broad range of legal subjects.

Specifically, for NJE, PAE and CPA with higher levels of difficulty, DISC-LawLLM surpasses all LLMs by a large margin on NJE and PAE. Especially for multi-answer questions that require more discerning judgment and reasoning, our model achieves an improvement of over 50% compared to the top-performing GPT-3.5-Turbo on NJE and PAE. This indicates that DISC-LawLLM has strong jurisprudential reasoning capabilities. Furthermore, the reason that generic LLMs occasionally outperform legal LLMs is likely to be the lack of few-shot instruction following ability during training.

6.2 Results in Subjective Evaluation

In the subjective evaluation, we utilize ChatGPT’s comprehension to evaluate the model’s performance on short answer questions against the Ground Truth. From Table 4, we can see that DISC-LawLLM achieves the best performance on most metrics. Compared to Chatlaw (Cui et al., 2023a), DISC-LawLLM shows a 6% increase in average performance. We can conclude that: 1) Leveraging the high-quality DISC-Law-SFT dataset enables DISC-LawLLM to generate more reliable responses, leading to outstanding scores in both ACC and CPL. 2) Through the deliberate cultivation of the model’s juridical thinking, the responses from DISC-LawLLM exhibit superior jurisprudential logic.

7 Applications

Our intelligent legal system, DISC-LawLLM, can serve various users across diverse scenarios. In this section, we showcase its application examples in three scenarios: legal professional tools, legal consultation, and examination assistant. The corre-

sponding figures are displayed in appendix A.

Legal Professional Tools. Our DISC-LawLLM simplifies the work of legal professionals by offering advanced tools for extracting legal elements, detecting legal events, analyzing legal cases, matching similar cases, generating judicial summaries, etc. Figure 5 shows two cases of DISC-LawLLM in legal event detection and legal summarization. In the first case, we see that DISC-LawLLM can extract the event trigger words and corresponding event types. In the second case, DISC-LawLLM can generate an accurate summary of the judicial case. These tools not only streamline judicial event monitoring and accelerate the decision-making process, but also facilitate other intelligent legal tasks.

Legal Consultation. Our DISC-LawLLM can offer legal consultation for dispute resolution, which greatly facilitates access to legal services and remote legal counseling for the general public. Figure 6 shows two cases of DISC-LawLLM for legal consultations about claims and debts and agreement drafting. In the first case, DISC-LawLLM effectively leverages facts in relevant legal base to provide a reliable response about debt apportionment. In the second example, DISC-LawLLM can offer precise drafting suggestions. These instances demonstrate DISC-LawLLM’s sound legal knowledge and reasoning proficiency.

Examination Assistant. For law students, our DISC-LawLLM serves as a tutor, helping to consolidate legal knowledge and providing solutions to exam questions. Figure 7 shows two cases of DISC-LawLLM in legal examinations and legal knowledge solutions. For the first case, our DISC-LawLLM can first predict the correct answer, and simultaneously unpack the rationale for answering this question. For the second case, our DISC-LawLLM can provide detailed legal explanations. Such ability is very valuable for law students, as DISC-LawLLM can help them gain deeper insights and dissect the answers to previously challenging questions, thereby improving their command of the subject.

In addition, we also show two examples with retrieval results in Figure 8, which can be extended for more application scenarios. Overall, our DISC-LawLLM can bridge the gap between LLMs and various judicial scenarios, which satisfies the demands of broad populations and has significant application value.

8 Conclusion

In this paper, we introduce DISC-LawLLM, an intelligent legal system for offering various legal services. Based on public NLP legal task datasets, legal raw text and open-source instruction datasets, we utilize ChatGPT to reconstruct legal responses following legal syllogism for supervised fine-tuning. To enhance the reliability of output responses, we incorporate an external retrieval module into our system. Through learning legal reasoning and knowledge retrieval capability, DISC-LawLLM outperforms existing legal LLMs on our constructed legal benchmark. This benchmark provides a comprehensive assessment from both subjective and objective perspectives. We release the constructed datasets and model weights to foster further research.

References

- Baichuan-inc. 2023. Baichuan-13b. <https://github.com/baichuan-inc/Baichuan-13B>.
- MJ Bommarito, Daniel Martin Katz, and E Detterman. 2018. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. *Research Handbook on Big Data Law*.
- CAIL. 2020. Cail2020. <https://github.com/china-ai-law-challenge/CAIL2020>.
- CAIL. 2022. Cail2022. <https://github.com/china-ai-law-challenge/CAIL2022>.
- Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1561–1571, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023a. Chatlaw: Open-source legal large language model with integrated external knowledge bases.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2023b. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.
- 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, Heng Wang, and Zhiyuan Liu. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *Chinese Computational Linguistics*, pages 439–451, Cham. Springer International Publishing.
- Anne von der Lieth Gardner. 1987. *An artificial intelligence approach to legal reasoning*. MIT press.
- Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *ArXiv*, abs/2305.15062.
- IDEA-CCNL. 2021. Fengshenbang-lm. <https://github.com/IDEA-CCNL/Fengshenbang-LM>.
- Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuan-Jing Huang. 2018. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3703–3714.
- Lu Ji, Zhongyu Wei, Jing Li, Qi Zhang, and Xuan-jing Huang. 2019. Discrete argument representation learning for interactive argument pair identification. *arXiv preprint arXiv:1911.01621*.
- Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. *arXiv preprint arXiv:2307.08321*.
- Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998.
- Haitao Li. 2023. Lexilaw. <https://github.com/CSHaitao/LexiLaw>.
- Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023. Lawgpt. https://github.com/LiuHC0428/LAW_GPT.
- 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.
- Meta. 2023. Llama. <https://github.com/facebookresearch/llama>.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. *Crosslingual generalization through multitask finetuning*.
- OpenAI. 2022. *Chatgpt: Optimizing language models for dialogue*.
- OpenAI. 2023. *Gpt-4 technical report*.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.
- Richard A Posner. 1990. *The problems of jurisprudence*. Harvard University Press.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.
- Pengxiao Song. 2023. Lawgpt. <https://github.com/pengxiao-song/LaWGPT>.
- Yun Song and Zhongyu Wei. 2021. Inferring association between alcohol addiction and defendant’s emotion based on sound at court. *Frontiers in Psychology*, 12:669780.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hananeh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.
- Chaojun Xiao, Haoxi 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*. *CoRR*, abs/1807.02478.
- Jianxin Yang. 2023. Firefly. <https://github.com/yangjianxin1/Firefly>.
- Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. *arXiv preprint arXiv:1905.03969*.
- 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, Dublin, Ireland. Association for Computational Linguistics.
- Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. *Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.
- ymcui. 2023. Chinese-llama-alpaca-2. <https://github.com/ymcui/Chinese-LLaMA-Alpaca-2>.

Jian Yuan, Zhongyu Wei, Yixu Gao, Wei Chen, Yun Song, Donghua Zhao, Jinglei Ma, Zhen Hu, Shaokun Zou, Donghai Li, et al. 2021. Overview of smp-cail2020-argmine: The interactive argument-pair extraction in judgement document challenge. *Data Intelligence*, 3(2):287–307.

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, Wenhao Huang, and Jie Fu. 2023. [Chinese open instruction generalist: A preliminary release](#).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#).

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. Jecqa: A legal-domain question answering dataset. In *Proceedings of AAAI*.

A Appendix: Supplementary Figures





Legal Professional Tools	
	<p>Input: 以下是一起司法案件:</p> <p>2014年5月11日13时许,在文安县兴隆宫镇大郭庄村某甲家中,被告人张某甲因家庭琐事与家人产生矛盾,将家中用色拉油塑料瓶盛放的汽油向房屋地面泼洒后试图点燃。在此过程中,其儿子张某乙因不满家人屡次争吵,从张某甲手中抢过盛放汽油的塑料瓶继续向屋内泼洒汽油后,用随身携带的打火机点燃后离开现场,张某甲未采取任何措施并将大门锁上后离开。之后,张某甲的父亲等人及时赶到将火扑灭,屋内沙发、被子等物被烧毁。2014年5月11日,被告人张某乙自动到文安县公安局投案。</p> <p>请识别出该法律案件中的所有事件触发词,以及其所对应的事件类型。例如,事件触发词'盗走'对应事件类型'盗窃财物'。请用JSON数组输出结果,数组中的每个元素也是一个数组,形如:[事件触发词,事件类型]。</p>
	<p>DISC-LawLLM: ["产生", "冲突", "矛盾", "冲突", "点燃", "放火", "争吵", "言语冲突", "抢", "抢夺财物", "点燃", "放火", "离开", "逃匿", "烧毁", "毁坏财物", "投案", "投案"]</p>
	<p>Input: 以下是一段涉法舆情文本:</p> <p>原标题: 内蒙古一幼教虐待多名儿童获刑1年半禁止从业5年图为宣判现场。检察院供图呼伦贝尔11月28日电(记者张林虎)记者28日从内蒙古自治区扎兰屯市人民检察院获悉,由该院提起公诉的王某某涉嫌虐待幼儿案一审公开宣判,被告人王某某被判有期徒刑1年半,并被“从业禁止”。公诉机关指控,被告人王某某身为对未成年人负有特殊职责的幼儿园教师,利用幼儿午睡时间,多次对其监护、管理的多名幼儿实施虐待行为,严重侵害了被害人的人格尊严和心理健康,其行为触犯了相关法律,犯罪事实清楚,证据确实充分,应该以虐待被看护人罪追究其刑事责任,并提出“从业禁止”的量刑建议。最终,法院作出一审判决,以虐待被看护人罪判处被告人王某某有期徒刑1年6个月,并针对虐待被看护人犯罪首次作出“从业禁止”,禁止其在刑罚执行完毕之日起5年内从事未成年人教育及相关职业。</p> <p>请对其进行摘要。</p>
	<p>DISC-LawLLM: 内蒙古扎兰屯市人民法院审理一起幼教虐待多名儿童案件,被告人王某某因虐待被看护人罪被判处有期徒刑一年六个月,同时还被处以五年内不得从事未成年人的教育工作及与之相关的职业。</p>

Figure 5: Case of our DISC-LawLLM as legal professional tools





Legal Consultation	
	<p>Input: 夫妻一方在婚姻关系存续期间以个人名义对外负债,所借款项没有用于夫妻共同生活的,债权人能否要求夫妻双方连带偿还?</p>
	<p>DISC-LawLLM: 根据《民法典》第一千零六十四条规定,如果夫妻一方在婚姻关系存续期间以个人名义对外负债,但该债务未被认定为夫妻共同债务的,则由举债方自行承担。然而,根据同一法条的规定,若债权人能够证明该债务用于了夫妻共同生活、共同生产经营或者基于夫妻双方共同意思表示的情况下,那么该债务将视为夫妻共同债务,应当由夫妻双方连带偿还。因此,如果债权人能提供证据证明该债务确实用于夫妻共同生活或共同生产经营,并且另一方未能提出相反证据予以反驳的话,债权人可以要求夫妻双方连带偿还该笔债务。</p>
	<p>Input: 离婚财产分割协议怎么写?</p>
	<p>DISC-LawLLM: 离婚财产分割协议书范本</p> <p>甲方: _____, 男, 汉族, ____年____月____日出生, 住址: ____。乙方: _____, 女, 汉族, ____年____月____日出生, 住址: ____。双方于____年____月____日在____区人民政府办理结婚登记手续, 婚后未生育子女。因双方性格不合无法共同生活, 夫妻感情已完全破裂, 现双方就自愿离婚一事达成如下协议: 1. 男女双方自愿离婚; 2. 财产处理: (1) 存款: 双方名下现有银行存款共____元, 双方各分一半, 为____元。分配方案为: 各自名下的存款保持不变, 但男方/女方应于____年____月____日前一次性支付____元给女方/男方。(2) 房屋: 夫妻共同所有的位于XXX的房地产所有权归女方所有, 房地产权证的业主姓名变更的手续自离婚后一个月内办理, 男方必须协助女方办理完毕变更的一切手续, 过户费用由女方负责。女方应一次性补偿男方____元的差额;(3) 其他财产: 婚前双方各自的财产归各自所有, 男女双方各自的私人生活用品及首饰归各自所有(附清单)。3. 债务处理: 双方确认在婚姻关系存续期间没有发生任何共同债务, 任何一方如对外负有债务的, 由负债方自行承担。4. 子女抚养: 年____月____日生育有一女儿, 取名为: _____, 离婚后女儿随____方生活, 方每月____日内支付____元作女儿的生活费直至女儿十八周岁时止, 教育费、医疗费等费用由双方平均分担, 非抚养方享有探望权, 另一方应予以配合。5. 违约责任的约定: 任何一方不按本协议约定履行义务的, 应付违约金____元给对方。本协议一式三份, 自婚姻登记机关颁发《离婚证》之日起生效, 男、乙双方各执一份, 婚姻登记机关存档一份。甲方: _____ 乙方: _____</p> <p>____日期: ____年____月____日</p>

Figure 6: Case of our DISC-LawLLM in legal consultation





Examination Assistant	
	<p>Input:以下是一道单项选择题, 请通过分析得到答案: 法院可以受理被害人提起的下列哪一附带民事诉讼案件? A.抢夺案, 要求被告人赔偿被夺走并变卖的手机 B.寻衅滋事案, 要求被告人赔偿所造成的物质损失 C.虐待被监管人案, 要求被告人赔偿因体罚虐待致身体损害所产生的医疗费 D.非法搜查案, 要求被告人赔偿因非法搜查所导致的物质损失</p>
	<p>DISC-LawLLM : 答案为B.寻衅滋事案, 要求被告人赔偿所造成的物质损失 解析: 本题考查刑事附带民事诉讼。根据《刑事诉讼法》第101条规定:“被害人因人身权利受到犯罪侵犯或者财物被犯罪分子毁坏而遭受物质损失的, 有权在刑事诉讼过程中提起附带民事诉讼;被害人死亡或者丧失行为能力的, 其法定代理人、近亲属有权提起附带民事诉讼。如果是国家财产、集体财产遭受损失的, 人民检察院在提起公诉的时候, 可以提起附带民事诉讼。选项B属于由于被告人的犯罪行为而受物质损失。选项B正确。根据《刑事诉讼法解释》第139条规定, 非法占有处置被害人财产的犯罪应当追缴或者退赔, 不能提起附带民事赔偿。选项A不当选。选项C、D都属于国家机关工作人员在行使职权时, 侵犯他人人身、财产权利构成犯罪, 属于国家赔偿的范围。C、D两项不当选。</p>
	<p>Input:管制与拘役的区别有哪些</p>
	<p>DISC-LawLLM :管制与拘役的区别有: 1、性质不同。管制的刑期较短, 但对犯罪分子的人身自由限制较大;而拘役则是对罪犯短期剥夺人身自由并关押改造的一种刑罚方法。2、执行场所不同。被判处拘役的犯罪分子在判决后由公安机关就近在看守所或者其他监管场所执行;而被判处管制的犯罪分子则在其居住地执行。3、期限不同。管制的期限为三个月以上两年以下;拘役的期限一般为一个月到六个月, 数罪并罚时不得超过一年。</p>

Figure 7: Case of our DISC-LawLLM as examination assistant

Retrieval Augmented DISC-LawLLM	
	<p>Input:王女士在某公司工作已满四年, 在怀孕五个月时告知公司人事并申请公司配合办理生育保险申请等事项, 孕期工作表现均符合公司公示制度要求。两周后, 公司借口以王女士无法胜任工作为由要与其解除劳动合同, 王女士如何维护自己的合法权益?</p>
	<p>DISC-LawLLM : 根据《劳动法》的规定, 王女士在怀孕五个月时告知公司并申请了公司配合办理生育保险申请等事项, 孕期工作表现均符合公司公示制度要求, 属于合法的。根据《女职工劳动保护特别规定》的规定, 用人单位不得因女职工怀孕、生育、哺乳降低其工资、予以辞退、与其解除劳动合同或者聘用合同, 因此公司不得以王女士无法胜任工作为由与其解除劳动合同。因此, 如果公司以此为理由要王女士解除劳动合同, 王女士有权维护自己的合法权益, 包括要求公司按照法定程序办理生育保险申请等事项, 并可向劳动仲裁机构提起劳动仲裁, 要求公司支付相应的经济补偿或其他合法赔偿。</p> <p>检索结果: 出处[1]《劳动法(2018-12-29)》第二十九条: 劳动者有下列情形之一的, 用人单位不得依据本法第二十六条、第二十七条的规定解除劳动合同: (一) 患职业病或者因工负伤并被确认丧失或者部分丧失劳动能力的; (二) 患病或者负伤, 在规定的医疗期内的; (三) 女职工在孕期、产期、哺乳期内的..... 出处[2]《女职工劳动保护特别规定(2012-04-28)》第五条: 用人单位不得因女职工怀孕、生育哺乳降低其工资、予以辞退、与其解除劳动合同或者聘用合同。</p>
	<p>Input:累犯就是指犯法被判刑, 在刑罚结束后一段时期内又再次犯法的情况。那么, 出狱后五年内故意犯罪一定是累犯吗?</p>
	<p>DISC-LawLLM :根据《刑法》第六十五条和第六十六条的规定, 出狱后五年内再次犯法的犯罪分子不一定是累犯。根据第六十五条的规定, 犯罪分子在刑罚执行完毕或赦免后, 在五年内再次犯应当判处有期徒刑以上刑罚之罪的, 是被认定为累犯。但是, 根据第六十六条的规定, 危害国家安全犯罪、恐怖活动犯罪、黑社会性质的组织犯罪的犯罪分子, 在刑罚执行完毕或赦免后, 在任何时候再犯上述任一类罪的, 都被视为累犯。因此, 出狱后五年内再次犯法的犯罪分子不一定是累犯, 具体情况需要根据犯罪行为的具体性质和法律规定进行判断。</p> <p>检索结果: 出处[1]《刑法》第六十五条: 【一般累犯】被判处有期徒刑以上刑罚的犯罪分子, 刑罚执行完毕或者赦免以后, 在五年以内再犯应当判处有期徒刑以上刑罚之罪的, 是累犯, 应当从重处罚, 但是过失犯罪和不满十八周岁的人犯罪的除外。前款规定的期限, 对于被假释的犯罪分子, 从假释期满之日起计算。 出处[2]《刑法》第六十六条: 【特别累犯】危害国家安全犯罪、恐怖活动犯罪、黑社会性质的组织犯罪的犯罪分子, 在刑罚执行完毕或者赦免以后, 在任何时候再犯上述任一类罪的, 都以累犯论处。</p>

Figure 8: Cases of our Retrieval Augmented DISC-LawLLM.