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

Shengbin Yue\textsuperscript{1,2}, Wei Chen\textsuperscript{3}, Siyuan Wang\textsuperscript{4}, Bingxuan Li\textsuperscript{4}, Chenchen Shen\textsuperscript{4}, Shujun Liu\textsuperscript{4}, Yuxuan Zhou\textsuperscript{4}, Yao Xiao\textsuperscript{7}, Song Yun\textsuperscript{6}, Xuanjing Huang\textsuperscript{5}, Zhongyu Wei\textsuperscript{1,4

\textsuperscript{1}Research Institute of Intelligent Complex Systems, Fudan University, China
\textsuperscript{2}Shanghai Center For Mathematical Sciences, Fudan University, China
\textsuperscript{3}School of Software Engineering, Huazhong University of Science and Technology, China
\textsuperscript{4}School of Data Science, Fudan University, China
\textsuperscript{5}School of Computer Science, Fudan University, China
\textsuperscript{6}Rule of Law Institute, Northwest University of Political and Law, China
\textsuperscript{7}New York University Shanghai, China

\{sbyue23, bxli16, ccschen22, 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 \url{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 simplifies 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-
Intelligent legal systems with legal reasoning and knowledge for delivering legal interpretation and deep understanding and precise reference of legal knowledge are constantly evolving with new laws and regulations. To address this challenge, we present DISC-LawLLM, our large language model tailored for building intelligent legal systems from both objective and subjective dimensions.

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).

---

<table>
<thead>
<tr>
<th>Question: the tenderer and the contractor agreed in the bidding documents that the quality standard of the project was qualified. After winning the bidding, the contractor made another agreement with the winning bidder that the performance bond would not be returned if the project did not receive the “Luban Award”. Is this agreement valid?</th>
</tr>
</thead>
</table>

**DISC-LawLLM:**

- **Major Premise:** According to Article 788 of Civil Code, “A construction project contract is a contract ……”.
- **Minor Premise:** The standards of the “Luban Award” attached to this award are not national mandatory standards, ……..the obligations imposed on the contractor by such a commitment are already higher than those agreed in the bidding contract, and actually change the quality standards of the work agreed in the bidding documents.

**Conclusion:** Therefore, the agreement has already constituted a change in the substantive content of the awarded contract, and should be invalid.

---

With legal reasoning on top of a general domain Chinese LLM with 13B parameters ¹. 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 evaluation.

---

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.
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-alpaca-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. LexiLaw, 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 labeler are involved to
Table 1: Statistics of DISC-Law-SFT Dataset.

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

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 recently 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 $X_1$ as follows:

\[
\text{In the legal syllogism, the major premise is articles of law, the minor premise is the facts of the case, and the conclusion is the judgment of case.}
\]

\[
\text{Case: } X
\]

\[
\text{Let us use legal syllogism to think and output the judgment:}
\]

### 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 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.
Objective Perspective

Level: Hard (Both Single & Multiple Answer)
- National Judicial Examination (NJE)
- Patent Agent Examination (PAE)
- Certified Public Accountant Qualification Examination (CPA)

Level: Normal (Both Single & Multiple Answer)
- Unified National Graduate Entrance Examination (UNGEE)

Level: Easy (Only Single Answer)
- Public Institutions and Functionary Examination (PFE)
- Question Bank of Legal Basic Knowledge (LBK)

Accuracy: 0-100%

Subjective Perspective

Short answer question

Question | LLM answer | Ground Truth

Criteria 1: Accuracy
Criteria 2: Completeness
Criteria 3: Clarity

Score: 1-5

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 framework. 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.
<table>
<thead>
<tr>
<th>Model</th>
<th>Size</th>
<th>Hard</th>
<th>Normal</th>
<th>Easy</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NJE</td>
<td>PAE</td>
<td>CPA</td>
<td>UNGEE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S M</td>
<td>S M</td>
<td>S M</td>
<td>S M</td>
</tr>
<tr>
<td>ChatGLM</td>
<td>6B</td>
<td>31.66 1.08</td>
<td>27.97 2.90</td>
<td>37.06 13.33</td>
<td>39.69 20.69</td>
</tr>
<tr>
<td>Baichuan-Chat</td>
<td>13B</td>
<td>31.47 10.15</td>
<td>29.66 8.70</td>
<td>35.53 19.17</td>
<td>50.00 27.59</td>
</tr>
<tr>
<td>Chinese-alpaca2</td>
<td>13B</td>
<td>25.7 10.15</td>
<td>30.51 11.59</td>
<td>32.99 19.17</td>
<td>40.94 21.84</td>
</tr>
<tr>
<td>GPT-3.5-turbo</td>
<td>175B</td>
<td>36.5 10.58</td>
<td>37.29 17.03</td>
<td>42.13 21.67</td>
<td>51.25 28.74</td>
</tr>
<tr>
<td>LawGPT</td>
<td>7B</td>
<td>22.91 6.26</td>
<td>31.36 7.61</td>
<td>25.38 16.67</td>
<td>30.31 13.79</td>
</tr>
<tr>
<td>Lawyer LLaMA</td>
<td>13B</td>
<td>35.75 5.62</td>
<td>32.20 6.52</td>
<td>29.95 13.33</td>
<td>32.50 14.94</td>
</tr>
<tr>
<td>ChatLaw</td>
<td>13B</td>
<td>27.56 7.99</td>
<td>31.36 9.42</td>
<td>35.53 11.67</td>
<td>35.62 17.24</td>
</tr>
<tr>
<td>DISC-LawLLM</td>
<td>13B</td>
<td>42.09 19.87</td>
<td>40.68 18.48</td>
<td>39.59 19.17</td>
<td>50.94 25.29</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Level</th>
<th>S</th>
<th>M</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPA</td>
<td>Hard</td>
<td>197</td>
<td>120</td>
<td>317</td>
</tr>
<tr>
<td>NJE</td>
<td></td>
<td>537</td>
<td>463</td>
<td>1000</td>
</tr>
<tr>
<td>PAE</td>
<td></td>
<td>118</td>
<td>276</td>
<td>394</td>
</tr>
<tr>
<td>UNGEE</td>
<td>Normal</td>
<td>320</td>
<td>87</td>
<td>407</td>
</tr>
<tr>
<td>LBK</td>
<td>Easy</td>
<td>275</td>
<td></td>
<td>275</td>
</tr>
<tr>
<td>PFE</td>
<td></td>
<td>170</td>
<td></td>
<td>170</td>
</tr>
</tbody>
</table>

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
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.

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

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); 4) 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).

#### 6.2 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


Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023a. Chatlaw: Open-source legal large language model with integrated external knowledge bases.


Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu,


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.


以下是一起司法案件：2014年5月11日13时许，在文安县兴隆宫镇大郭庄村张某甲家中，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。

请识别出该法律案件中的所有事件触发词，并分别进行标注。
以下是一道单项选择题，请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

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、财产利益构成犯罪，属于国家赔偿的范围。C、D两项不当选。