LAiW: A Chinese Legal Large Language Models Benchmark

Anonymous ACL submission

Abstract

General and legal domain LLMs have demonstrated strong performance in various tasks of LegalAI. However, the current evaluations of these LLMs in LegalAI lack consistency with the legal logic, making LLMs difficult to understand and trust by legal experts. To address this challenge, we are the first to build the Chinese legal LLMs benchmark LAiW, based on the logic of legal syllogism. We categorize the legal capabilities of LLMs into three levels to align with the thinking process of legal experts and legal syllogism: basic information retrieval, legal foundation inference, and complex legal application. Each level collects and tailors multiple tasks to ensure a comprehensive evaluation. Through automatic evaluation of current general and legal domain LLMs on our benchmark, we indicate that although LLMs can answer complex legal questions, the LLMs do not possess the rigorous logical processes inherent in legal syllogism, which may pose obstacles to be accepted by legal experts. To further confirm this scenario of LLMs in legal application, we incorporate manual evaluation with legal experts. The results not only confirm the above conclusion but also reveal the important role of pretraining for LLMs in enhancing legal logic, which may improve the future development of the legal LLM.

1 Introduction

With the emergence of ChatGPT and GPT-4 and their excellent text processing capabilities (Zhao et al., 2023), many researchers have paid attention to the applications of large language models (LLMs) in various fields (Wang et al., 2023; Xie et al., 2023; Ko and Lee, 2023). In the field of legal artificial intelligence (LegalAI), which specifically studies how artificial intelligence can assist in legal construction (Zhong et al., 2020b; Locke and Zuccon, 2022; Feng et al., 2022), LLMs, especially those specializing in Chinese law, show strong capabilities in generating legal text (Cui et al., 2023a; Pengxiao et al., 2023; Wen and He, 2023).

However, due to the opaque nature of LLMs, legal experts are cautious about their practical application in the law (Dahl et al., 2024). They believe that the lack of understanding the logic and the thinking process of LLMs in legal practice may greatly impact the fairness of the law[1]. More importantly, the current Chinese legal LLMs and benchmarks have not fully explored this issue. Although current Chinese legal LLMs cover a wide range of legal tasks and utilize pre-training (Wen and He, 2023) or fine-tuning (Wu et al., 2023; Cui et al., 2023a) to acquire knowledge or capabilities in the legal field, they only focus on improving the effectiveness of these tasks without analyzing the relevance of the legal logic between the tasks. Existing benchmarks for evaluating these models are also constructed based on these tasks level(Yue et al., 2023; Fei et al., 2023). They focus on whether certain types of legal tasks, such as legal question answering and consultation(Zhong et al., 2020b; Choi, 2023; Steenhuis et al., 2023). Therefore, the classification and relationship of types/levels cannot demonstrate the logical application of LLMs in law. It is important to explore the abilities of LLMs from the logic of the legal application perspective, to ensure that legal experts have a better understanding of LLMs in legal tasks and trust them.

In the opinion of legal experts, the application of LLMs in law should adhere to the logic of legal practice framework, known as the legal syllogism, involving the acquisition of evidence, legal articles, conclusions, and their interconnections (Kuppa et al., 2023; Trozze et al., 2023), as shown in Table 1. Firstly, the ability to extract information from legal texts, then the ability to provide a reliable and reasoned answer based on solid legal

[1] https://github.com/liuchengyuan123/LegalLLMEvaluation/

knowledge, and ultimately the ability to form a complete response. This entire process avoids logical confusion and ensures the regularity of legal logic and the reliability of legal conclusions.

Major Premise: The relevant legal articles. ***Minor Premise***: The information and evidence pertinent to a case. ***Conclusion***: The judicial decision based on these premise.
For Example: In criminal law, when judging someone, we need to first find **relevant legal articles** based on **evidence**, then calculate the **judgment result** based on these articles, and provide a well-organized and logical **judgment text**.

Table 1: The Legal Syllogism.

In this work, to investigate the above-mentioned issue, we propose the first Chinese legal LLM benchmark LAiW[2] based on the logic of legal syllogism. In this benchmark, corresponding to the thought process of legal syllogism, we categorize the legal capabilities of LLMs into three levels, from simple to difficult: basic information retrieval (BIR), legal foundation inference (LFI), and complex legal application (CLA). Among them, BIR focuses on the general NLP capabilities of LLMs and some legal evidence, knowledge, and category determination, which are tailored for the Major Premise and Minor Premise in legal; LFI emphasizes the performance of LLMs in simple application tasks in the legal domain, which tries to let LLMs give a Conclusion based on Major Premise and Minor Premise; CLA focuses on the performance of LLMs in complex tasks in the legal domain, which requires support from the abilities developed in the first two levels and integrates them to form the entire legal logical process. Based on these capabilities, our benchmark collects and reconstruct 14 tasks from the existing LegalAI tasks.

When conducting benchmark evaluations, we performed both automatic evaluations and additional manual evaluations. For automatic evaluations, we not only evaluate existing Chinese legal LLMs but also focused on the base models of these Chinese legal LLMs and more effective general LLMs. The results of automatic evaluations indicate that while existing LLMs have strong text generation capabilities for complex legal applications, they are unable to meet the underlying logic in legal applications in basic information retrieval and legal foundation inference. This demonstrates that the powerful legal logical process of LLMs does not come from the step-by-step legal syllogism. Therefore, we conduct additional manual evaluations to specifically investigate the reasons behind this and to confirm the effectiveness of automatic evaluations. Through evaluations by legal experts, we find that in some complex legal applications with relatively lenient requirements for legal logic, LLMs' powerful generation ability cleverly bridges the gap in legal logic. However, in more demanding scenarios, they exhibit significant discrepancies from real results. We find that legal syllogism of LLMs may be learned from the pretrain stage, which is difficult to learn through fine-tuning alone. This provides insight for future practical improvements for LLMs in the legal field.

Our contributions are as follows:

- We are proud to introduce the first Chinese legal LLMs benchmark LAiW, which is designed based on the logic of legal syllogism. We categorize the legal capabilities of LLMs into three levels to facilitate a more precise evaluation of legal logic process of LLMs in legal practice and to enhance legal experts' understanding of LLMs.

- Based on our automatic evaluation, we demonstrate that the current legal LLMs do not have the logic of legal syllogism. While LLMs demonstrate strong text generation abilities to complete complex legal applications, struggling to achieve satisfactory performance in adhering to the basic legal logic framework.

- We invite legal experts for manual evaluations to further explore the reasons for the lack of legal syllogism in LLMs. This indicates the need of the tasks of the legal logic to pretrain LLMs for future development.

2 Related Work

Chinese Legal LLMs. We summarize the current Chinese legal LLMs and some general models in Table 4. Most of these Chinese legal LLMs focus on the ultimate application tasks in the legal field and are generally fine-tuned on some general LLMs. For instance, LawGPT_zh (Liu et al., 2023), Lawyer-LLaMA (Huang et al., 2023a), ChatLaw (Cui et al., 2023a), Fuzi-Mingcha (Wu et al., 2023), and LexiLaw developed the ability to answer legal questions and provide legal consultations by fine-tuning on related legal data. To compensate for the lack of legal knowledge due to only fine-tuning, these LLMs introduce additional legal knowledge

[2]It means "AI in LAW".

databases for retrieval to supplement. However, the accuracy and comprehensiveness of the knowledge base may be a major limiting factor for these LLMs. The other Chinese legal LLMs adopted the pretraining or continued pretraining to enhance the legal knowledge of LLMs, such as LaWGPT (Pengxiao et al., 2023), wisdomInterrogatory, and HanFei (Wen and He, 2023). They collect a large amount of legal text data, covering a wider range of legal tasks such as element extraction and case classification. These have a noticeable impact on improving the overall effectiveness of LLMs in legal applications. However, the Chinese legal LLMs mentioned above mainly focus on the performance of legal application in each tasks, which rarely consider whether they meet the logical requirements of legal practice. It is important to evaluate their legal logic, which is of utmost concern to legal experts.

Legal LLMs Benchmark. The development of LegalAI has led to a substantial quantity of tasks that combine law and computer science, from NLP-focused legal NER and legal text summarization (Kanapala et al., 2019) to legal-focused similar case matching (Locke and Zuccon, 2022; Sansone and Sperlí, 2022), providing ample data for evaluating Chinese legal LLMs (Zhong et al., 2020b). When categorizing from a legal perspective, it also encompasses the logic of the entire legal process from the legal elements extraction (Cao et al., 2022; Zhang et al., 2022a; Zhong et al., 2020a) to legal judgment prediction (Feng et al., 2022; Cui et al., 2023b). Based on these tasks, LawBench (Fei et al., 2023) built an automatic evaluation framework for Chinese legal LLMs, which concerns the memorization, understanding, and application of legal knowledge. DISC-Law-Eval Benchmark (Yue et al., 2023) also based on the aforementioned tasks divides the evaluation into objective and subjective parts. The objective section assesses knowledge retention and reasoning abilities in the legal examination, and the subjective part uses GPT-3.5 Turbo to score the accuracy, completeness, and clarity of the answers. These frameworks have helped us understand the capabilities of legal LLMs from the perspective of knowledge systems. However, whether these LLMs can be accepted by legal experts from legal logic is still a question worthy of evaluation. In this work, we focus on addressing this issue from the legal syllogism.

3 Benchmark Construction

In this section, we divide LLMs' abilities levels based on the Legal Syllogism in practice, and construct our Chinese legal LLMs benchmark LAiW based on these levels. To ensure comprehensive evaluation, we incorporate both automatic evaluation using computable metrics and manual evaluation by legal experts.

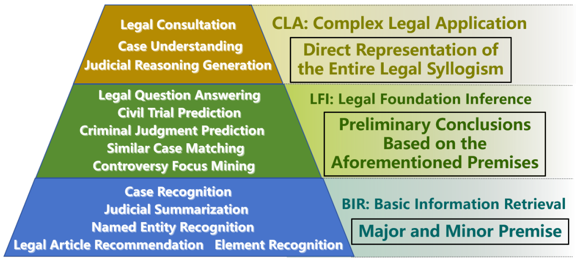

Figure 1: Multi-level Legal Capabilities of LLMs.

3.1 The Logic of Legal Practice for LLMs

In contemporary legal practice, the logical framework is based on Syllogism (Wróblewski, 1974; Patterson, 2013). It typically consists of three parts: the major premise, the minor premise and the conclusion, which is derived from the major and minor premises. As shown in Table 1, in legal practice, this entails assessing the information and evidence pertinent to a case (minor premise), identifying the relevant legal articles (major premise), and reaching a judicial decision based on these factors (conclusion). This systematic approach underscores the intricate interplay between legal articles and factual circumstances in legal practice.

To ensure that LLMs also have the aforementioned logical framework and remain synchronized with legal practice, we should also divide the abilities of LLMs into the aforementioned logical stages with 14 tasks. Specifically, we categorize the legal abilities of LLMs into three levels and try to align them with the logic of legal syllogism, as illustrated in Figure 1. By merging the process of acquiring minor premise and major premise, we construct the capability level of **basic information retrieval**. Building upon this foundation, we develop the capability level of **legal foundation inference** to draw preliminary conclusions based on the minor and major premises. Additionally, to assess the direct representation of the entire legal syllogism, we have created the capability level of **complex legal application**[3].

[3] Appendix A.2 provides more details for each tasks.

Capability	Task	Primary Origin Dataset	LAiW	Domain	Task Type	Class	Balance
BIR	Legal Article Recommendation	CAIL2018 (Xiao et al., 2018)	1000	Criminal	Classification	3	0.231
	Element Recognition	CAIL-2019 (Zhang et al., 2022a)	1000	Civil	Classification	20	0.002
	Named Entity Recognition	CAIL-2021 (Cao et al., 2022)	1040	Criminal	Named Entity Recognition	-	-
	Judicial Summarization	CAIL-2020 (Huang et al., 2023b)	364	Civil	Text Generation	-	-
	Case Recognition	CJRC (Duan et al., 2019)	2000	Criminal, Civil	Classification	2	0.499
LFI	Controversy Focus Mining	LAIC-2021	306	-	Classification	10	0.029
	Similar Case Matching	CAIL-2019 (Xiao et al., 2019)	260	Civil	Classification	2	0.450
	Charge Prediction	Criminal-S (Hu et al., 2018)	827	Criminal	Classification	3	0.172
	Prison Term Prediction	MLMN (Ge et al., 2021)	349	Criminal	Classification	3	0.074
	Civil Trial Prediction	MSJudeg (Ma et al., 2021)	800	Civil	Classification	3	0.065
	Legal Question Answering	JEC-QA (Zhong et al., 2020c)	855	-	Classification	4	0.201
CLA	Judicial Reasoning Generation	AC-NLG (Wu et al., 2020)	834	Civil	Text Generation	-	-
	Case Understanding	CJRC (Duan et al., 2019)	1054	Criminal, Civil	Text Generation	-	-
	Legal Consultation	CrimeKgAssitant (Liu et al., 2023)	916	-	Text Generation	-	-

Table 2: Statistical information of our dataset. All datasets are sourced from open-source. In classification tasks, "Balance" refers to the proportion of the least represented class in the dataset compared to the total dataset size. It can be observed that the dataset labels for the four tasks: Element Recognition, Controversy Focus Mining, Prison Term Prediction, and Civil Trial Prediction, are significantly unbalanced.

3.1.1 BIR: Basic Information Retrieval

We design the Basic Information Retrieval level with 5 tasks to assess the fundamental abilities of LLMs in legal logic, corresponding to directly accessible text information, minor premises, and major premises, such as legal evidence, legal knowledge, and category determination. This is the most fundamental step within the framework of legal syllogism, identifying all the necessary elements for its following reasoning.

Specifically, we first consider three tasks that are well-defined in the fields of law and NLP: Named Entity Recognition, Judicial Summarization, and Case Recognition. They identify and summarize the key elements in legal texts, and classify cases as either Criminal or Civil. Although these tasks may not require extensive legal knowledge from LLMs, they can yield a wealth of foundational information useful for both legal and computational purposes from the text.

We also consider two other tasks in the legal domain, namely Legal Article Recommendation and Element Recognition. The first task is to catch the major premises by finding relevant legal articles. The second task is to catch minor premises by identifying their relevant elements.

3.1.2 LFI: Legal Foundation Inference

The Legal Foundation Inference level follows Syllogism's idea to explore the ability of LLMs to derive basic results and some judgment conclusions from minor premises and major premises. This also constitutes the core step in legal syllogism, as it connects all the parts within legal syllogism.

We can divide 6 tasks for this capability into three parts. The first part presents the basic results of some simple legal applications, including Controversial Focus Mining and Similar Case Matching. Controversial Focus Mining is an intermediate result obtained in civil law based on the underlying circumstances and legal articles, used to determine the core issues of concern for both the plaintiff and the defendant. Similar Case Matching involves finding similar cases based on the current case situation and referring to these cases to ensure the fairness of the judgment. The second part involves predicting the outcomes of the court judgment conclusion. Since criminal law and civil law are two main branches of law, we have 3 tasks. Charge Prediction and Prison Term Prediction for criminal law, Civil Trial Prediction for civil law. Finally, The third part involves another application task, Legal Question Answering, that requires some fundamental integrated capabilities and focuses on the simple application of legal knowledge. Based on the information provided, LLMs provide some basic legal responses.

3.1.3 CLA: Complex Legal Application

For this capability, we endeavor to integrate the steps of legal syllogism mentioned above, exploring whether LLMs can effectively accomplish responses based on legal syllogism in tasks. We consider 3 challenging tasks that LLMs may be required to complete a complex legal reasoning and application task: Judicial Reasoning Generation, Case Understanding, and Legal Consultation. Judicial Reasoning Generation involves the complete reproduction of the logical process from major and minor premises to conclusions in legal judgments. Case Understanding, on the other hand, analyzes the logic from the perspective of understanding, from major and minor premises to conclusion. Legal Consultation utilizes this logic from the perspective of a legal professional to provide assistance.

3.2 Datasets Construction

With the mentioned criteria for capability division and task preparation, we construct the evaluation dataset for our LAiW benchmark based on the majority of open-source datasets and a small amount of proprietary data. The dataset is divided into two parts: Automatic and Manual.

3.2.1 Automatic Evaluation Datasets

To facilitate a more efficient evaluation of LLMs, we construct all 14 tasks mentioned above into datasets that can be automatically assessed shown in Table 2. The primary sources of this data include previous years' CAIL competition data (Xiao et al., 2018; Zhang et al., 2022a; Huang et al., 2023b), as well as the most commonly used open-source data (Ge et al., 2021; Wu et al., 2020; Liu et al., 2023). We cover a wide range of legal areas, including criminal law, civil law, constitutional law, social law, and economic law, to encompass as many legal scenarios as possible..

During the construction of the dataset, we designed different prompts for various tasks to ensure LLMs can provide related answers. We validated the quality of prompts using ChatGPT and confirmed their validity through legal experts. Currently, all tasks exist in a zero-shot format[4].

3.2.2 Manual Evaluation Datasets

As shown in automatic evaluation results 5.2, we observed that these LLMs may not align with the logic of legal syllogism. LLMs seem to be able to directly acquire complex legal application capabilities but perform poorly in following the syllogism framework. To further investigate the reasons behind the phenomenon caused by LLMs, we add a manual evaluation focus on the third level.

Due to the cost of the manual evaluation, we focus on two tasks that are more oriented toward LLMs for logical reasoning: Judicial Reasoning Generation and Legal Consultation. These two tasks are respectively directly or indirectly related to the syllogism framework.[5]

4 Evaluation for Benchmark

In this section, we provide the criteria, the metrics and scoring method for automatic evaluation and manual evaluation.

4.1 Automatic Evaluation

Automatic Evaluation Legal Tasks contains classification tasks, named entity recognition tasks, and text generation tasks. Table 3 presents the evaluation metrics[6] for each task.

Task	Metric
Classification	Acc, F1, Miss, Mcc
Named Entity Recognition	Entity-Acc
text generation	ROUGE-1, ROUGE-2, ROUGE-L

Table 3: The metrics for automatic evaluation.

To evaluate the overall legal capabilities of LLMs, we further select a few key indicators for each task and calculate legal scores for LLMs based on these indicators as shown in Equation (1).

$$\begin{cases} S_{\text{classification}} = F1 * 100, \\ S_{\text{text generation}} = \frac{1}{3}(R1 + R2 + RL) * 100, \\ S_{\text{named entity recognition}} = \text{Entity-Acc} * 100. \end{cases} \tag{1}$$

Subsequently, the total score is calculated by averaging the scores of the three levels, which in turn are determined by averaging the scores of tasks within each level.

4.2 Manual Evaluation

First, to ensure the reliability of the assessment, we present criteria with several legal experts for manual evaluation, shown in Table 5[7].

We adopt the approach used in studies (Dubois et al., 2023; Li et al., 2023) for manual evaluation, considering legal experts as evaluators, using

[4]Examples and the detailed processing methods can be found in Appendix A.2 and Appendix A.3.

[5]The detailed processing methods for the datasets are outlined in Appendix A.4.

[6]The details of these metrics are provided in Appendix C.

[7]A more detailed description about these criteria is provided in Appendix B.2.

Model	Model Size	Model Domain	From	Baseline	Creater	URL
GPT-4 (OpenAI, 2023)	-	General	Api	-	OpenAI	[1]
ChatGPT	-	General	Api	-	OpenAI	[2]
Baichuan2-Chat (Baichuan, 2023)	13B	General	Open	-	Baichuan Inc	[3]
Baichuan	7B	General	Open	-	Baichuan Inc	[4]
ChatGLM (Du et al., 2022)	6B	General	Open	-	Tsinghua, Zhipu	[5]
Llama (Touvron et al., 2023a)	7B	General	Application	-	Meta AI	[6]
Llama (Touvron et al., 2023a)	13B	General	Application	-	Meta AI	[6]
Llama2-Chat (Touvron et al., 2023b)	7B	General	Application	-	Meta AI	[7]
Chinese-LLaMA (Cui et al., 2023c)	7B	General	Open	Llama-7B	Yiming Cui	[8]
Chinese-LLaMA (Cui et al., 2023c)	13B	General	Open	Llama-13B	Yiming Cui	[8]
Ziya-LLaMA(Zhang et al., 2022b)	13B	General	Open	Llama-13B	IDEA-CCNL	[9]
HanFei (Wen and He, 2023)	7B	Law	Open	-	SIAT NLP	[10]
wisdomInterrogatory	7B	Law	Open	Baichuan-7B	ZJU, Alibaba, e.t	[11]
Fuzi-Mingcha (Wu et al., 2023)	6B	Law	Open	ChatGLM-6B	irlab-sdu	[12]
LexiLaw	6B	Law	Open	ChatGLM-6B	Haitao Li	[13]
LaWGPT (Pengxiao et al., 2023)	7B	Law	Open	Chinese-LLaMA-7B	Pengxiao Song	[14]
Lawyer-LLaMA (Huang et al., 2023a)	13B	Law	Open	Chinese-LLaMA-13B	Quzhe Huang	[15]
ChatLaw (Cui et al., 2023a)	13B	Law	Open	Ziya-LLaMA-13B	PKU-YUAN's Group	[16]

Table 4: The LLMs evaluated in our work. LaWGPT and wisdomInterrogatory undergo pre-training on Chinese-LLaMA and Baichuan respectively, followed by fine-tuning. HanFei does not have a baseline model. Apart from GPT-4 and ChatGPT, these general LLMs have a parameter size of 7-13B to ensure a size similar to legal LLMs.

Task	Criteria
Judicial Reasoning Generation	Completeness, Relevance, Accuracy
Legal Consultation	Fluency, Relevance, Comprehensibility

Table 5: The assessment criteria for manual evaluation.

reference answers as the baseline to calculate the win rate for the target LLMs. For example, when using the reference answer as the baseline, legal experts comprehensively assess the output of the target LLM and the reference answer from multiple judgment dimensions, and then choose the most satisfactory response.

5 Experiment

In this section, we present relevant experiment settings and highlight the key results of the Legal Syllogism in LLMs.

5.1 Experiment Settings

For the automatic evaluation, We evaluate 18 LLMs, including 7 mainstream legal LLMs (Cui et al., 2023a; Pengxiao et al., 2023), their corresponding 6 baseline LLMs (Du et al., 2022; Cui et al., 2023c; Zhang et al., 2022a), and 5 more effective general LLMs (Baichuan, 2023; Touvron et al., 2023a) such as GPT-4 and ChatGPT. For fairness in evaluation, all LLMs did not utilize legal knowledge databases. Table 4 lists more detailed information about these LLMs.

For the manual evaluation, We choose the top-performing four legal LLMs in our automatic evaluation. They are Fuz-Mingcha (Wu et al., 2023), HanFei (Wen and He, 2023), Lawyer-LLaMa (Huang et al., 2023a), and LexiLaw. Furthermore, we also conducted manual assessments of the performance of both GPT-4 and ChatGPT.

5.2 Automatic Evaluation Results

The scores for each level and the total score of our automatic evaluation are shown in Table 6[8]. We analyze the results from two different aspects: overall results and the legal logic of Chinese Legal LLMs.

Overall results. When compared to GPT-4 and ChatGPT, there still exists a significant gap between the current open-source LLMs and specifically trained legal LLMs.

From Table 6, we find that GPT-4 and ChatGPT maintain optimal performance in most tasks. They significantly outperform the current open-source LLMs at various levels of scoring. Among the open-source LLMs, only Baichuan2-Chat, ChatGLM, and Ziya-LLaMA achieve a total score of 45 or above. However, their performance in the BIR and LFI levels still lags far behind GPT-4 and Chat-

[8]The complete results of each task are available in Appendix D.1.

Model	Basic Information Retrieval						Legal Foundation Inference							Complex Legal Application				Total Score
	B_1	B_2	B_3	B_4	B_5	Avg.	L_1	L_2	L_3	L_4	L_5	L_6	Avg.	C_1	C_2	C_3	Avg.	
GPT-4	99.20	82.27	80.67	42.72	99.75	80.92	80.50	45.94	100.00	65.58	70.43	53.14	69.27	37.22	96.19	42.66	58.69	**69.63**
ChatGPT	99.05	79.32	61.73	41.01	98.85	75.99	57.16	46.17	99.28	47.35	62.85	37.08	58.32	35.64	90.70	47.55	57.96	**64.09**
Baichuan2-13B-Chat	45.07	52.18	47.31	26.67	97.14	53.67	4.12	2.99	17.50	61.43	67.91	38.24	32.03	52.61	81.29	41.31	58.40	**48.04**
Baichuan-7B	17.81	2.87	0.00	26.89	58.45	21.20	1.74	0.00	1.18	1.03	64.50	24.32	15.46	40.27	33.79	18.51	30.86	22.51
ChatGLM-6B	72.55	49.82	1.06	42.87	91.27	51.51	14.18	39.03	67.57	44.84	33.02	23.86	37.08	35.39	86.90	35.02	52.44	**47.01**
Llama-7B	19.53	1.43	0.00	11.40	23.23	11.12	1.31	0.00	35.19	1.03	49.15	5.74	15.40	0.61	56.08	10.93	22.54	16.35
Llama-13B	28.16	7.66	0.00	9.94	46.80	18.51	1.86	0.00	36.79	5.80	40.46	5.57	15.08	11.19	65.68	11.34	29.40	21.00
Llama2-7B-Chat	48.24	11.93	0.19	15.79	83.17	31.86	0.74	0.00	3.88	7.31	62.09	2.59	12.77	28.76	69.51	17.65	38.64	27.76
Chinese-LLaMA-7B	24.39	7.45	0.00	30.77	48.97	22.32	2.02	0.76	31.79	1.03	65.24	8.63	18.25	26.34	62.31	13.81	34.16	24.91
Chinese-LLaMA-13B	30.34	5.47	0.00	7.73	61.56	21.02	3.28	5.05	20.21	5.33	64.46	16.60	19.16	18.86	73.15	12.40	34.80	24.99
Ziya-LLaMA-13B	66.39	58.42	48.94	38.85	94.73	61.47	5.64	0.76	53.18	55.62	36.07	25.38	29.44	30.12	83.96	25.26	46.45	**45.79**
HanFei-7B	24.91	7.25	51.63	21.14	82.18	37.42	1.15	0.00	5.27	2.73	66.81	22.03	16.33	51.31	81.19	27.43	53.31	35.69
wisdomInterrogatory-7B	0.39	0.19	0.00	34.75	27.99	12.66	3.57	35.38	2.32	1.30	16.76	3.34	10.45	13.91	68.02	18.17	33.37	18.83
Fuzi-Mingcha-6B	58.95	12.58	0.38	47.92	78.57	39.68	4.70	20.84	31.53	48.40	32.66	26.64	27.46	49.55	80.48	34.10	54.71	40.62
LexiLaw-6B	47.16	2.89	31.35	41.79	83.43	41.32	2.11	18.49	3.40	6.42	4.35	18.51	8.88	25.85	80.81	24.52	43.73	31.31
LaWGPT-7B	10.15	2.59	0.00	27.69	36.92	15.47	1.62	0.00	20.04	1.03	54.55	8.40	14.27	35.23	65.62	14.11	38.32	22.69
Lawyer-LLaMA-13B	20.26	1.52	7.88	51.13	73.44	30.85	2.19	0.76	0.24	2.12	12.75	20.26	6.39	34.00	85.68	31.83	50.50	29.25
ChatLaw-13B	67.08	31.29	52.21	41.33	98.20	58.02	0.00	0.00	37.82	30.85	6.58	0.00	12.54	0.00	20.23	0.00	6.74	25.77

Table 6: The all scores of LLMs at various levels of the LAiW based on equation (1). We use bold to indicate the top-performing five LLMs overall. Here, B_1 to B_5 respectively represent the tasks: Legal Article Recommendation, Element Recognition, Named Entity Recognition, Judicial Summarization, and Case Recognition. L_1 to L_6 respectively represent the tasks: Controversy Focus Mining, Similar Case Matching, Charge Prediction, Prison Term Prediction, Civil Trial Prediction, and Legal Question Answering. C_1 to C_3 respectively represent the tasks: Judicial Reasoning Generation, Case Understanding, and Legal Consultation.

GPT. As for the specifically trained legal LLMs, the top four performing ones are Fuzi-Mingcha, HanFei, LexiLaw, and Lawyer-LLaMA. However, their overall scores are lower, all below 45.

We believe that the reason for this phenomenon is twofold: first, due to the large number of parameters in GPT-4 and ChatGPT; second, we during the pretraining phase, GPT-4 and ChatGPT may have been exposed to a larger amount of data. Since the open-source LLMs we selected are primarily aimed at the Chinese community, the data they collect may be more limited compared to GPT-4 and ChatGPT. GPT-4 and ChatGPT cover a wide range of legal data in multiple languages. In this case, it is reasonable for them to have higher scores in the BIR and LFI levels which focus on the basic legal logic and legal knowledge.

The Legal Syllogism in Chinese Legal LLMs. Most of the Legal LLMs cannot follow the Legal Syllogism framework. While they demonstrate strong text generation abilities in complex legal applications, they perform poorly in other tasks.

Observing Table 6, it is evident that the majority of legal LLMs score nearly 20 points higher in the application of direct logic (CLA level) compared to the scores in BIR and LFI levels. This is contrary to the logic typically found in law. It suggests that these LLMs seem to have learned the patterns of generating legal texts directly, but have not grasped the legal reasoning behind these patterns. As a result, LLMs are unable to effectively identify the major and minor premises in law and lack the ability to reason to a conclusion. However, for the BIR level, ChatLaw stands out among legal LLMs. It instead has a strong ability at the BIR level, which may stem from the outstanding performance of its base model Ziya - LLaMA at this level.

This raises concerns that current legal LLMs may not meet the expectations of legal experts. The performance demonstrated by LLMs shows a very weak correlation with the logical framework of the law, potentially jeopardizing trust in LLMs within the legal domain.

5.3 Manual Evaluation Results

Model	Judicial Reasoning Generation			Legal Consultation		
	Total Score	Win Rate	Std	Total Score	Win Rate	Std
GPT-4	44.72	0.38	0.18	43.97	**0.85**	0.15
ChatGPT	41.74	0.35	0.27	**48.79**	0.79	0.12
Fuzi-Mingcha	**63.58**	**0.65**	0.35	35.22	0.51	0.19
HanFei	60.13	0.59	0.26	27.06	0.33	0.06
LexiLaw	43.48	0.31	0.15	25.53	0.24	0.02
Lawyer-LLaMA	39.61	0.30	0.26	33.27	0.51	0.21

Table 7: The average win rate (WR) of LLMs for the Judicial Reasoning Generation and Legal Consultation tasks. The total score represents the score obtained by LLMs through automatic evaluation on our benchmark. We use bold to indicate the best and underline to indicate the second-best.

According to the assessment criteria for expert evaluation in Section 4.2, and the calculated average win rate scores of three legal experts shown in Table 7[9]. Based on these results, we have three findings.

Manual evaluation and automatic evaluation share similarities. This enhances the reliability of our automatic evaluation. From Table 7, we can observe that the results of manual evaluation and automatic evaluation are similar. For instance, in both evaluation rounds, Fuzi-Mingcha and HanFei performed best in the Judicial Reasoning Generation task, while GPT-4 and ChatGPT excelled in the Legal Consultation task. In addition, despite its shortcomings as an automatic evaluation metric in many cases, Rouge still demonstrates a certain level of capability when reflecting legal logic. This indicates that our automatic evaluation can provide a reliable path for the legal logic assessment of legal LLMs and further reduce manual effort. Additionally, our assessment of legal logic is granular, and the degree of emphasis on legal logic in different scenarios can also be reflected by our automatic evaluation of different tasks.

The lack of Legal Syllogism in LLMs still exists in Complex Legal Applications. For the task of Judicial Reasoning Generation that requires a strong understanding of legal logic, even models with powerful text generation capabilities like GPT-4 and ChatGPT may have deficiencies in legal logic. As described in Section 4.2, the Judicial Reasoning Generation task focuses on accuracy, such as the correct citation of legal articles and reasoning based on the citation. This directly connects to the basic logic of legal. Therefore, most of the LLMs' win rates are much lower than 0.5, indicating that strong text generation capabilities cannot directly replace legal logic.

For tasks like Legal Consultation, there is a lower requirement for legal logic but a higher requirement for fluency. Therefore, during the manual evaluation, legal experts tend to prefer models with stronger language capabilities, which is the strength of GPT-4 and ChatGPT. This capability can also be learned by legal LLMs through instruction tuning. As a result, the final evaluation results of legal experts also reflect this, giving higher win rates to all LLMs, with most even surpassing the annotated answers.

The future of Chinese Legal LLMs. Fine-tuned legal LLMs have enhanced the normativity of legal text generation, but they may sacrifice legal logic. Furthermore, for legal LLMs, undergoing additional pre-training on legal text could be the pathway to acquiring diverse legal capabilities and understanding the logic of legal syllogism.

From manual evaluation, legal experts find that legal LLMs such as Fuzi-Mingcha, WisdomInterrogatory, LaWGPT and Lawyer-LLaMA have the powerful normativity of generated texts in legal text generation. Referring to Table 6, we can further find that the acquisition of this normativity may stem from fine-tuning LLMs on CLA-level tasks compared to their base models. This enables LLMs to respond in a certain style, albeit not within the logical framework of Legal Syllogism. Moreover, such fine-tuning may result in a decline in performance at the BIR and LFI levels.

On the other hand, legal LLMs like HanFei, which focus more on pre-training, may indicate how Chinese Legal LLMs acquire ability and logic. HanFei, although it is based on an older LLM structure (Bloomz), with extensive pre-training on legal texts, it demonstrates capabilities on par with subsequent legal LLMs from automatic and manual evaluation. Furthermore, GPT-4 and ChatGPT, models with extensive pre-training on large corpora, also showed excellent performance at the BIR and LFI levels. These findings indicate that developing legal reasoning and comprehensive abilities may require learning from a significant amount of pre-training text, rather than just fine-tuning.

6 Conclusion

This paper aims to construct a Chinese Legal LLMs benchmark based on the logic of legal syllogism in practice. To match the process of syllogism in legal logic step by step, the benchmark categorizes LLM legal capabilities into three levels and encompasses 14 tasks. During benchmark evaluations, automatic and manual evaluations were conducted. Automatic results showed that existing LLMs excel in text generation for complex legal applications but struggle with basic information retrieval and legal foundation inference, leading to a lack of legal logic and distrust among legal experts. Manual evaluations revealed that while LLMs may bridge the gap in legal logic in some application scenarios, they still exhibit significant discrepancies as legal experts. This underscores the necessity for further pretraining of LLMs in the legal domain to gain the logic of legal syllogism rather than solely relying on fine-tuning.

[9]The detailed win rate scores and agreements results are available in Appendix D.2, Appendix D.3 and Appendix D.4.

7 Limitations and Future Work

Due to the significant amount of work required to construct this benchmark and complete the evaluation, we also acknowledge the following two limitations and areas for future work:

1) In the manual evaluation experiment, to save workload, only a portion of the data and LLMs are sampled and chosen for evaluation, rather than assessing all of them. In the future, we will collaborate more with legal experts to ensure a more comprehensive human assessment.

2) Most of the tasks are collected and reconstructed from publicly available legal data, which may not comprehensively evaluate the logic of legal practice for LLMs. We will further develop additional tasks to refine the logic of legal practice at each stage.

8 Ethics Statement

Due to the sensitivity of the legal field, we have conducted a comprehensive review of the relevant data in this benchmark. The open-source datasets we used all have corresponding licenses. We have masked sensitive information, such as names, phone numbers, and IDs, and legal experts have conducted ethical evaluations.

Acknowledgements

References

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yu Cao, Yuanyuan Sun, Ce Xu, Chunnan Li, Jinming Du, and Hongfei Lin. 2022. Cailie 1.0: A dataset for challenge of ai in law-information extraction v1. 0. *AI Open*, 3:208–212.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13.

Jonathan H Choi. 2023. How to use large language models for empirical legal research. *Journal of Institutional and Theoretical Economics (Forthcoming)*.

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

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023b. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023c. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*.

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

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

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction: A survey of the state of the art. *IJCAI. ijcai. org*, pages 5461–9.

Jidong Ge, Yunyun Huang, Xiaoyu Shen, Chuanyi Li, and Wei Hu. 2021. Learning fine-grained fact-article correspondence in legal cases. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3694–3706.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023a. Lawyer llama technical report. *ArXiv*, abs/2305.15062.

Yue Huang, Lijuan Sun, Chong Han, and Jian Guo. 2023b. A high-precision two-stage legal judgment summarization. *Mathematics*, 11(6):1320.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51:371–402.

Hyungjin Ko and Jaewook Lee. 2023. Can chatgpt improve investment decision? from a portfolio management perspective. *From a Portfolio Management Perspective*.

Aditya Kuppa, Nikon Rasumov-Rahe, and Marc Voses. 2023. Chain of reference prompting helps llm to think like a lawyer. In *Generative AI+ Law Workshop*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023. Lawgpt: 中文法律对话语言模型. https://github.com/LiuHC0428/LAW_GPT.

Daniel Locke and Guido Zuccon. 2022. Case law retrieval: problems, methods, challenges and evaluations in the last 20 years. *arXiv preprint arXiv:2202.07209*.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal judgment prediction with multi-stage case representation learning in the real court setting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–1002.

OpenAI. 2023. Gpt-4 technical report.

Edwin W Patterson. 2013. Logic in the law. In *Logic, Probability, and Presumptions in Legal Reasoning*, pages 287–321. Routledge.

Song Pengxiao, Zhou Zhi, and cainiao. 2023. Lawgpt: 基于中文法律知识的大语言模型. https://github.com/pengxiao-song/LaWGPT.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Quinten Steenhuis, David Colarusso, and Bryce Willey. 2023. Weaving pathways for justice with gpt: Llm-driven automated drafting of interactive legal applications. *arXiv preprint arXiv:2312.09198*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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

Arianna Trozze, Toby Davies, and Bennett Kleinberg. 2023. Large language models in cryptocurrency securities cases: Can chatgpt replace lawyers? *arXiv preprint arXiv:2308.06032*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Jibao Wen and Wanwei He. 2023. Hanfei. https://github.com/siat-nlp/HanFei.

Jerzy Wróblewski. 1974. Legal syllogism and rationality of judicial decision. *Rechtstheorie*, 5:33.

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

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780.

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

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

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

Dian Zhang, Hewei Zhang, Long Wang, Jiamei Cui, Wen Zheng, et al. 2022a. Recognition of chinese legal elements based on transfer learning and semantic relevance. *Wireless Communications and Mobile Computing*, 2022.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022b. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1250–1257.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. 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. 2020c. Jecqa: a legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708.

A More Details of Data Construction

A.1 Data Source

For the convenience of researchers, Table 8 lists the original sources of our reconstructed dataset.

A.2 Automatic Evaluation Dataset

In this section, we provide the construction details for the LAiW datasets of each task.

A.2.1 BIR: Basic Information Retrieval

Legal Article Recommendation.

Definition: Legal Article Recommendation aims to provide relevant articles based on the description of the case.

Description: It comes from the first stage data of the CAIL-2018, aimed at providing relevant legal articles based on case descriptions. We selected the top three legal articles with their corresponding charges, namely the crime of dangerous driving, theft, and intentional injury. The three charges correspond to Article 133, Article 264, and Article 234 of the Criminal Law of the People’s Republic of China.

Prompt: "Based on the relevant description provided below, predict the applicable law article. The options are (’133’, ’264’, ’234’). Your answer must be one of these three articles. These articles represent the legal provisions in the Criminal Law of the People’s Republic of China. Among them, Article ’133’ refers to ’Violating regulations on transportation management, resulting in a major accident causing serious injury, death, or significant loss of public or private property’. Article ’264’ refers to ’Stealing public or private property, or committing theft multiple times, burglary, armed theft, or pickpocketing’. Article ’234’ refers to ’Intentionally causing bodily harm to others’. Text:"

"请根据下面给定的案件的相关描述预测其涉及的法条，可供选择的法条为(’133’, ’264’, ’234’)，回答只能是这三个法条中的一个。这三个法条代表《中华人民共和国刑法》中的法律条文，其中，法条’133’表示’违反交通运输管理法规，因而发生重大事故，致人重伤、死亡或者使公私财产遭受重大损失’，法条’264’表示’盗窃公私财物，或者多次盗窃、入户盗窃、携带凶器盗窃、扒窃’，法条’234’表示’故意伤害他人身体’。文本:"

Element Recognition.

Definition: Element Recognition analyzes and assesses each sentence to identify the pivotal elements of the case.

Description: It comes from the element recognition track of the CAIL-2019, aiming to automatically extract key factual descriptions from case descriptions. The original dataset primarily involves marriage, labor disputes, and loan disputes. We selected the labor dispute dataset.

Prompt: "Based on the partial paragraphs of the arbitral awards in the field of labor disputes below, identify the elements involved. The selectable elements are (’LB1’, ’LB2’, ’LB3’, ’LB4’, ’LB5’, ’LB6’, ’LB7’, ’LB8’, ’LB9’, ’LB10’, ’LB11’, ’LB12’, ’LB13’, ’LB14’, ’LB15’, ’LB16’, ’LB17’, ’LB18’, ’LB19’, ’LB20’). The options are as follows: ’LB1’ represents ’termination of labor relations’, ’LB2’ represents ’payment of wages’, ’LB3’ represents ’payment of economic compensation’, ’LB4’ represents ’non-payment of full labor remuneration’, ’LB5’ represents ’existence of labor relations’, ’LB6’ represents ’no labor contract signed’, ’LB7’ represents ’labor con-

Dataset	URL
CAIL-2018	http://cail.cipsc.org.cn/task_summit.html?raceID=1&cail_tag=2018
CAIL-2019	https://github.com/china-ai-law-challenge/CAIL2019
CAIL-2021	https://github.com/isLouisHsu/CAIL2021-information-extraction/tree/master
CAIL-2020	http://cail.cipsc.org.cn/task_summit.html?raceID=4&cail_tag=2022
CJRC	https://github.com/china-ai-law-challenge/CAIL2019/tree/master
LAIC-2021	https://laic.cjbdi.com/
Criminal-S	https://github.com/thunlp/attribute_charge
MLMN	https://github.com/gjdnju/MLMN
MSJudge	https://github.com/mly-nlp/LJP-MSJudge
JEC-QA	https://jecqa.thunlp.org/
AC-NLG	https://github.com/wuyiquan/AC-NLG
CrimeKgAsitant	https://github.com/LiuHC0428/LAW-GPT

Table 8: The original source of the datasets utilized in the experiment. We conducted extensive cleaning and reconstruction on these data to align their format with legal logic, in order to obtain instruction datasets for evaluation.

tract signed', 'LB8' represents 'payment of overtime wages', 'LB9' represents 'payment of double wages compensation for unsigned labor contracts', 'LB10' represents 'payment of work-related injury compensation', 'LB11' represents 'not raised at the labor arbitration stage', 'LB12' represents 'non-payment of compensation for illegal termination of labor relations', 'LB13' represents 'economic layoffs', 'LB14' represents 'non-payment of bonuses', 'LB15' represents 'illegally collecting property from workers', 'LB16' represents 'specialized occupations', 'LB17' represents 'payment of work-related death allowance|funeral allowance|bereavement allowance', 'LB18' represents 'advance notice of termination by the employer', 'LB19' represents 'corporate legal status has ceased', 'LB20' represents 'mediation agreement exists'. Text:"

"请根据以下劳动争议领域的裁判文书的部分句段，识别其涉及的要素，可供选择的要素有('LB1', 'LB2', 'LB3', 'LB4', 'LB5', 'LB6', 'LB7', 'LB8', 'LB9', 'LB10','LB11', 'LB12', 'LB13', 'LB14', 'LB15', 'LB16', 'LB17', 'LB18', 'LB19', 'LB20')，回答只能是这二十个选项中的一个。这二十个选项中，'LB1'表示'解除劳动关系'，'LB2'表示'支付工资'，'LB3'表示'支付经济补偿金'，'LB4'表示'未支付足额劳动报酬'，'LB5'表示'存在劳动关系'，'LB6'表示'未签订劳动合同'，'LB7'表示'签订劳动合同'，'LB8'表示'支付加班工资'，'LB9'表示'支付未签订劳动合同二倍工资赔偿'，'LB10'表示'支付工伤赔偿'，'LB11'表示'劳动仲裁阶段未提起'，'LB12'表示'不支付违法解除劳动关系赔偿金'，'LB13'表示'经济性裁员'，'LB14'表示'不支付奖金'，'LB15'表示'违法向劳动者收取财物'，'LB16'表示'特殊工种'，'LB17'表示'支付工亡补助金|丧葬补助金|抚恤金'，'LB18'表示'用人单位提前通知解除'，'LB19'表示'法人资格已灭失'，'LB20'表示'有调解协议'。文本："

Named Entity Recognition.

Definition: Named Entity Recognition aims to extract nouns and phrases with legal characteristics from various legal documents.

Description: It comes from the Information Extraction competition of CAIL-2021, aiming to extract the main content of judgments. The original dataset covers 10 legal entities, including "criminal suspect," "victim," etc. We selected five entities: "criminal suspect," "victim," "time," "stolen items," and "item value." We filtered out samples with non-nested entities. We used five prompts, each corresponding to one of the five legal entities.

Prompt: "Your task is to extract the entity 'suspect' from the text below. If this entity does not exist, the answer is 'No'. Text: "A set of stolen 'Jingqiu' brand batteries worth 1488 yuan." Answer:"

"你的任务是从下面的文本中提取'犯罪嫌疑人'实体，如果不存在这个实体，则回答'No'。文本：被盗"京球"牌蓄电池一组价值人民币1488元。回答："

Judicial Summarization.

Definition: Judicial Summarization aims to condense, summarize, and synthesize the content of legal documents.

Description: It comes from the Judicial Summary competition of CAIL-2020, aiming to extract the main content of judgments. We removed certain information from the original text of each sample, including case number, case title, judges, trial time, etc., as we believe this information has little impact on the quality of summary generation. Additionally, we only kept samples with a text length less than 1.5k.

Prompt: "Please extract an abstract from the legal document given below and express its main content in shorter, more coherent and natural words. text:"

"请对下面给的这篇法律文书提取摘要，用更短、更连贯、更自然的文字表达其主要内容。文本："

Case Recognition.

Definition: Case Recognition aims to determine, based on the relevant description of the case, whether it pertains to a criminal or civil matter.

Description: It comes from CJRC, aiming to determine whether a given case is a criminal or civil case based on relevant case descriptions. We sampled criminal and civil cases in nearly a 1:1 ratio.

Prompt: " Please determine whether the following case belongs to criminal or civil cases based on the title or relevant description text, and your response should be one of the two options. Text:"

"请根据以下案件的标题或者相关描述文本，判断该案件属于刑事案件还是民事案件，并且你的回答应该只能是其中一个。文本："

A.2.2 LFI: Legal Foundation Inference

Controversy Focus Mining.

Definition: Controversial Focus Mining aims to extract the logical and interactive arguments between the defense and prosecution in legal documents, which will be analyzed as a key component for the tasks that relate to the case result.

Description: It comes from the Controversy Focus Recognition task of LAIC, aiming to identify and detect the disputed focal points based on the original plaintiff's claims and defense contents in legal judgments. We selected samples that meet the following conditions: 1) contain only one disputed focal point, 2) have a text length less than 3k, and 3) involve the top ten disputed focal points in terms of frequency. Consequently, we restructured the dataset into a classification task, where the model is required to correctly identify the disputed focal point from the ten available options for each sample.

Prompt: "Please select the most appropriate dispute focus based on the plaintiff's claims and defendant's defense in the judgment document. The options are ('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J'), representing ten dispute focuses respectively. You only need to return the letter of the correct option. Among them, 'A' represents 'determination of the amount of engineering funds', 'B' represents 'determination of the amount of damages compensation', 'C' represents 'dispute over principal/loan agreement/written agreement or electronic agreement/expressions of borrowing intention', 'D' represents 'dispute over principal/loan agreement/written agreement or electronic agreement/principal amount', 'E' represents 'liability determination', 'F' represents 'whether there is a breakdown of relationship', 'G' represents 'guarantee liability/claim for warranty', 'H' represents 'existence of labor relations', 'I' represents 'contractual effectiveness issue', 'J' represents 'responsibility assumption'. Text:"

"请根据裁判文书中原被告的诉请及答辩内容选择出一个最匹配的争议焦点。可供选择的回答为('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J')，这十个选项分别代表十个争议焦点，你只需要返回正确选项的字母。其中，回答'A'表示'工程款数额认定'，回答'B'表示'损失赔偿数额认定'，回答'C'表示'本金争议/借贷合意/书面协议or电子协议/借款的意思表示'，回答'D'表示'本金争议/借贷合意/书面协议or电子协议/本金（金额）'，回答'E'表示'责任认定'，回答'F'表示'感情是否破裂'，回答'G'表示'担保责任/保证责任诉求'，回答'H'表示'是否存在劳动关系'，回答'I'表示'合同效力问题'，回答'J'表示'责任承担'。文本："

Similar Case Matching.

Definition: Similar Case Matching aims to find cases that bear the closest resemblance, which is a core aspect of various legal systems worldwide, as they require consistent judgments for similar cases to ensure the fairness of the law.

Description: It comes from CAIL2019-SCM, which aims to match similar cases based on factual descriptions. Each entry in the original dataset contains three fields labeled 'A,' 'B,' and 'C,' representing three legal factual descriptions. Our task is to determine, given three legal documents A, B, and C, which one (B or C) is more similar to A. Additionally, each selected case has a length not exceeding 2k.

Prompt: "Based on the content of Case A, select the case that is more similar to Case A. The options are ('B', 'C'). The length of the answer is limited to 3 characters, meaning you only need to provide the letter of the correct option. 'B' indicates that Case B is more similar to Case A, while 'C' indicates that Case C is more similar to Case A."

"请根据案件A的内容，选择与案件A相似度更高的案件，可供选择的回答为（'B', 'C'），回答的文本长度限制为3个字符，即只回答正确选项的字母。其中，回答'B'表示案件B与案件A相似度更高，回答'C'表示案件C与案件A相似度更高。"

Charge Prediction.

Definition: It is the sub-task of Criminal Judgment Prediction task. Criminal Judgment Prediction involves predicting the guilt or innocence of the defendant, along with the potential sentencing, based on the results of basic legal NLP, including the facts of the case, the evidence presented, and the applicable law articles.

Description: It is from the Criminal-S dataset, which consists of criminal cases published by CJO. As each case is well-structured and divided into multiple sections such as facts, court opinions, and judgment results, the authors of this dataset chose the facts section of each case as input and selected 149 different charges as output. In this paper, we specifically chose the charges of "Theft," "Intentional Smuggling," and "Drug Trafficking, Selling, Transporting, and Manufacturing" as our focus. Each sample corresponds to a unique charge.

Prompt: "Based on the given description of the case below, predict the crime it involves. The options are ('69', '50', '124'). You can only choose one of these three options. '69' represents 'theft', '50' represents 'intentional injury', and '124' represents 'smuggling, selling, transporting, or manufacturing drugs'. Text:"

"请根据下面给定的案件的相关描述预测其涉及的罪名，可供选择的回答为('69', '50', '124')，回答只能是这三个选项中的一个。这三个选项代表了三个罪名，其中，罪名'69'表示'盗窃罪'，罪名'50'表示'故意伤害罪'，罪名'124'表示'走私、贩卖、运输、制造毒品罪'。文本："

Prison Term Prediction.

Definition: It is the sub-task of Criminal Judgment Prediction task, which is defined in Charge Prediction task.

Description: It comes from MLMN, aiming to learn fine-grained correspondences of factual-Articles in legal cases. The original dataset is divided into crimes of injury and traffic accidents. Based on the original data's months of imprisonment, the labels are categorized into five classes. In this paper, we further categorized the sentences into three classes: the first class includes non-punishment and detention, the second class includes imprisonment of less than 1 year and 1 year to less than 3 years, and the third class includes imprisonment of 3 years to less than 10 years.

Prompt: "Based on the given description of the case below, predict the possible sentence the defendant may receive. The options are ('A', 'B', 'C'). You can only choose one of these three options. 'A' represents 'non-criminal punishment' or 'detention', 'B' represents 'fixed-term imprisonment of less than 3 years', and 'C' represents 'fixed-term imprisonment of 3 years or more but less than 10 years'. Text:"

"请根据下面给定的案件的相关描述预测被告人可能被判的刑期，可供选择的回答为('A', 'B', 'C')，回答只能是这三个选项中的一个。这三个选项对应了三个刑期区间，其中，回答'A'表示'免予刑事处罚'或'拘役'，回答'B'表示'3年以下有期徒刑'，回答'C'表示'3年及3年以上，10年以下有期徒刑'。文本："

Civil Trial Prediction.

Definition: Civil Trial Prediction task involves using factual descriptions to predict the judgment of the defendant in response to the plaintiff's claim, which we should consider the Controversial Focus.

Description: It comes from MSJudge, aiming to predict opinions on each claim based on case-related descriptions and claims. The original dataset includes court factual descriptions, multiple claims, and judgments for each claim. We extracted samples with only a unique claim and sampled them based on the distribution of judgment results.

Prompt: "Based on the factual description of the civil case provided below and a litigation request, provide an overall judgment prediction for the litigation request. Your response can only be one of the three options ('A', 'B', 'C'). 'A' indicates support for the litigation request, 'B' indicates partial support for the litigation request, and 'C' indicates opposition to the litigation request."

"根据下面给定民事案件的事实描述和一个诉讼请求，给出你对该诉讼请求的一个整体

裁判预测，你的回答只能是(’A’, ’B’, ’C’)三个选项中的一个。其中，’A’表示支持诉讼请求，’B’表示部分支持诉讼请求，’C’表示反对诉讼请求。"

Legal Question Answering.

Definition: Legal Question Answering utilizes the model’s legal knowledge to address the national judicial examination, which encompasses various specific legal types.

Description: It is from a question-answering dataset collected from the China National Judicial Examination, which includes both single-choice and multiple-choice questions. The goal is to predict answers using the presented legal questions and relevant articles. We selected only the single-choice questions for our analysis.

Prompt: "Please answer the question based on the judicial examination question below. There is only one correct answer among the options (’A’, ’B’, ’C’, ’D’). You don’t need to provide a detailed analysis of the question, just select the correct answer."

"请根据下面的司法考试题目回答问题，选项(’A’, ’B’, ‘C’, ’D’)中只有一个正确答案。你不需要返回对题目的具体分析，只需选出正确的答案。"

A.2.3 CLA: Complex Legal Application

Judicial Reasoning Generation.

Definition: Judicial Reasoning Generation aims to generate relevant legal reasoning texts based on the factual description of the case. It is a complex reasoning task, because the court requires further elaboration on the reasoning behind the judgment based on the determination of the facts of the case. This task also involves aligning with the logical structure of syllogism in law.

Description: It comes from the AC-NLG dataset, constructed from private lending cases, which are the most common category in civil cases. The focus is on the task of generating court opinions in civil cases. This task takes the plaintiff’s claims and factual descriptions as input and generates the corresponding court opinions as output.

Prompt: "Please generate corresponding "the court holds that" content based on the "litigation requests" and "trial findings" provided in the brackets below."

"请你根据下面中括号里的’诉讼请求’和’审理查明’内容生成对应的’本院认为’内容。"

Case Understanding.

Definition: Case Understanding is expected to provide reasonable and compliant answers based on the questions posed regarding the case-related descriptions in the judicial documents, which is also a complex reasoning task.

Description: It also comes from the CJRC dataset, which includes 10,000 documents and nearly 50,000 questions with answers. These documents are from judgment files, and the questions are annotated by legal experts. Each document contains multiple questions. In this paper, we selected only the training set from the original data, where each question has only one standard answer.

Prompt: "Based on the provided "legal text material" content, answer the corresponding "question" to complete the task of fragment extraction-based reading comprehension. Specifically, you need to correctly answer the "question", and the answer is limited to a clause (or fragment) from the "legal text material". Please provide your answer in the format ”’Answer: A”’, where A represents the correct clause (or fragment) from the "legal text material"."

"请你根据下面提供的’法律文本材料’内容，回答相应的’问题’，以完成片段抽取式的阅读理解任务。具体来说，你要正确回答’问题’，并且答案限定是’法律文本材料’的一个子句（或片段）。请你以”’答案：A”’的格式给出回答，其中A表示’法律文本材料’中正确的子句（或片段）。"

Legal Consultation.

Definition: Legal Consultation covers a wide range of legal areas and aims to provide accurate, clear, and reliable answers based on the legal questions provided by the different users. Therefore, it usually requires the sum of the aforementioned capabilities to provide professional and reliable analysis.

Description: It comes from the CrimeKgAssistant dataset, where ChatGPT has been utilized to rephrase answers based on the Q&A pairs from CrimeKgAssistant. The goal is to generate answers that are more detailed and linguistically well-organized compared to the original responses. We further filtered question-answer pairs by identifying responses containing phrases like "抱歉" or "无法准确回答", and cases where questions contained numerous "?" symbols or were linguistically awkward.

Prompt: "If you are a lawyer, please answer the legal consultation question below based on the real scenario."

"假设你是一名律师，请回答下面这个真实情景下的中文法律咨询问题。"

A.3 Some Automatic Evaluation Examples

This is an appendix for examples of our three tasks: classification, named entity recognition and text generation. Then we respectively choose Case Recognition, Named Entity Recognition and Legal Consultation tasks as example prompts for these three tasks.

In named entity recognition task, the {entity} include: 犯罪嫌疑人,受害人, 时间, 物品价值, 被盗物品, totaling five entities.

A.3.1 Example for Classification Tasks

Please determine whether the following case belongs to criminal or civil cases based on the title or relevant description text, and your response should be one of the two options.

Text: "The People's Procuratorate of Neixiang County accuses that, from June 26 to June 29, 2016, the defendant, Zhang, organized personnel to cut down a large number of poplar trees on the roadside farmland in Shangwangzhuangzu, Miaobei Village, Chimei Town, without obtaining a timber harvesting permit. According to the appraisal by the Neixiang County Forestry Investigation and Design Team, a total of 128 poplar trees were felled, with a total living wood volume of 32.5521 cubic meters. On June 29, 2016, the defendant, Zhang, voluntarily surrendered to the Neixiang County Forest Public Security Bureau."

Answer:

请根据以下案件的标题或者相关描述文本，判断该案件属于刑事案件还是民事案件，并且你的回答应该只能是其中一个。

文本：'内乡县人民检察院指控，2016年6月26日至6月29日期间，被告人张某在未办理林木采伐许可证的情况下，组织人员将其购买的位于赤眉镇庙北村上王庄组路边耕地里的大量杨树砍伐。经内乡县林业调查设计队鉴定，共砍伐杨树128株，计活立木蓄积32.5521立方米。2016年6月29日，被告人张某主动到内乡县森林公安局投案自首。'

回答:

A.3.2 Example for Named Entity Recognized Tasks

Your task is to extract the '{entity}' entity from the text below. If this entity does not exist, the answer is 'No'.

Text: "A set of stolen 'Jingqiu' brand batteries worth 1488 yuan."

Answer:

你的任务是从下面的文本中提取'{entity}'实体，如果不存在这个实体，则回答'No'。

文本：被盗"京球"牌蓄电池一组价值人民币1488元。

回答:

A.3.3 Example for Text Generation Tasks

If you are a lawyer, please answer the legal consultation question below based on the real scenario.

'Question': I was driving straight ahead, and a tricycle coming from the opposite direction hit me as it came out of the gas station, causing injuries to people. Who is more responsible, and how is the responsibility divided?

假设你是一名律师，请回答下面这个真实情景下的中文法律咨询问题。

'问题'：我开车直行，对面三轮车从加油站出来撞了，人受伤了，是谁的责任大，怎么划分责任的?

A.4 Manual Evaluation Dataset

In this section, we focus on the manual evaluation of the Judicial Reasoning Generation and Legal Consultation tasks.

Legal Consultation. We directly use the legal evaluation dataset from the previous automatic evaluation of the Legal Consultation task, sampling 50 data points as the artificial evaluation dataset for the Legal Consultation task.

Judicial Reasoning Generation. We reconstructed the evaluation dataset. Our dataset is sourced from the China Judgements Online (CJO), where all are written judgment of first instance. We extract the sections in the documents related to the court identified that, claims, and court hold that. In the end, our reconstructed Judicial Reasoning Generation manual evaluation dataset consists of 50 data points, covering five charges: kidnapping, trafficking of women and children, fraud, robbery, and extortion, with 10 data points for each charge.

B More Details of Manual Evaluation

B.1 Data License

The Legal Consultation is sourced from a public dataset, while the Judicial Reasoning Generation comes from our private dataset. All personally identifiable information such as names, phone numbers,

Capability	Task	Metrics	GPT-4	ChatGPT	HanFei	wisdomInterrogatory	Fuzi-Mingcha	LexiLaw	LaWGPT	Lawyer-LLaMA	ChatLaw
BIR	Legal Article Recommendation	Acc	**0.9890**	0.9880	0.1690	0.0020	0.5540	0.5240	0.0590	0.1280	0.6570
		Miss	0.0060	0.0050	0.6530	0.9940	0.1840	0.0100	0.8770	0.7570	0.1000
		F1	**0.9920**	0.9905	0.2491	0.0039	0.5895	0.4716	0.1015	0.2026	0.6708
	Element Recognition	Acc	**0.8170**	0.7910	0.0600	0.0010	0.1390	0.0230	0.0480	0.0080	0.3050
		Miss	0	0.0010	0.7650	0.9970	0.0750	0.8250	0.2900	0.9700	0.2880
		F1	**0.8227**	0.7932	0.0725	0.0019	0.1258	0.0289	0.0259	0.0152	0.3129
		Mcc	0.7960	0.7656	0.0289	0.0110	0.0861	0.0113	-0.0108	0.0198	0.2381
	Named Entity Recognition	Entity-Acc	**0.8067**	0.6173	0.5163	0	0.0038	0.3135	0	0.0788	0.5221
	Judicial Summarization	ROUGE-1	0.5549	0.5463	0.2834	0.4592	0.6243	0.5406	0.3894	**0.6467**	0.5362
		ROUGE-2	0.2982	0.2849	0.1359	0.2400	0.3423	0.2947	0.1746	**0.3877**	0.3000
		ROUGE-L	0.4285	0.3990	0.2150	0.3433	0.4710	0.4184	0.2668	**0.4994**	0.4036
	Case Recognition	Acc	**0.9975**	0.9885	0.8270	0.2820	0.7935	0.8380	0.4670	0.7505	0.9815
		Miss	0	0	0	0.4435	0.0025	0.0010	0.1790	0.0005	0.0010
		F1	**0.9975**	0.9885	0.8218	0.2799	0.7857	0.8343	0.3692	0.7344	0.9820
LFI	Controversy Focus Mining	Acc	**0.8072**	0.5458	0.0229	0.0817	0.049	0.0359	0.0458	0.0392	0
		Miss	0.0196	0.0196	0.3595	0.2484	0.4085	0.6536	0.4641	0.4967	1
		F1	**0.8050**	0.5716	0.0115	0.0357	0.0470	0.0211	0.0162	0.0219	0
		Mcc	0.7662	0.4713	-0.0284	0.0393	0.0066	0.0210	0.0159	0.0079	0
	Similar Case Matching	Acc	**0.5692**	0.5500	0	0.3885	0.1654	0.1231	0	0.0038	0
		Miss	0	0.0038	0.9962	0.3423	0.6692	0.7769	1	0.9923	1
		F1	0.4594	**0.4617**	0	0.3538	0.2084	0.1849	0	0.0076	0
	Charge Prediction	Acc	**1**	0.9927	0.1717	0.0121	0.2044	0.0181	0.1330	0.0012	0.4631
		Miss	0	0	0.0060	0.9649	0.7352	0.9528	0.7509	0.9915	0.0278
		F1	**1**	0.9928	0.0527	0.0232	0.3153	0.0340	0.2004	0.0024	0.3782
	Prison Term Prediction	Acc	**0.6533**	0.4499	0.0802	0.0287	0.4097	0.0716	0.0745	0.0115	0.2579
		Miss	0	0	0	0.7450	0.2923	0.4900	0	0.9628	0.0573
		F1	**0.6558**	0.4735	0.0273	0.0130	0.484	0.0642	0.0103	0.0212	0.3085
		Mcc	0.3353	0.1705	-0.0125	0.0239	0.0810	-0.0226	0	0.0240	-0.0467
	Civil Trial Prediction	Acc	0.6775	0.5925	**0.7675**	0.0950	0.2183	0.0266	0.5038	0.0712	0.1500
		Miss	0.0525	0.0075	0.0025	0.8950	0.6713	0.9686	0.3425	0.8988	0.1138
		F1	**0.7043**	0.6285	0.6681	0.1676	0.3266	0.0435	0.5455	0.1275	0.0658
		Mcc	0.2657	0.1929	0.0155	0.0602	0.0165	-0.0046	0.0023	0.0051	0.0283
	Legal Question Answering	Acc	**0.5298**	0.3789	0.2398	0.0222	0.2456	0.2199	0.1731	0.2175	0
		Miss	0.0012	0	0.0538	0.8760	0.1871	0.0959	0.2094	0.2094	1
		F1	**0.5314**	0.3708	0.2203	0.0334	0.2664	0.1851	0.0840	0.2026	0
CLA	Judicial Reasoning Generation	ROUGE-1	0.5193	0.4985	**0.6882**	0.2105	0.6804	0.3613	0.4943	0.4809	-
		ROUGE-2	0.2473	0.238	**0.3723**	0.0698	0.3411	0.1517	0.2286	0.2091	-
		ROUGE-L	0.3499	0.3326	**0.4788**	0.1371	0.4651	0.2626	0.3340	0.3300	-
	Case Understanding	ROUGE-1	**0.9650**	0.9168	0.8219	0.7502	0.8173	0.8307	0.7187	0.8765	0.2061
		ROUGE-2	**0.9568**	0.8919	0.7917	0.5778	0.7837	0.7735	0.5625	0.8268	0.1962
		ROUGE-L	**0.9640**	0.9122	0.8220	0.7127	0.8134	0.8200	0.6873	0.8671	0.2047
	Legal Consultation	ROUGE-1	0.5974	**0.6482**	0.3777	0.2518	0.4797	0.3436	0.1956	0.4514	-
		ROUGE-2	0.2758	**0.3197**	0.1693	0.0980	0.2086	0.1391	0.0660	0.1992	-
		ROUGE-L	0.4066	**0.4585**	0.2759	0.1953	0.3346	0.2529	0.1617	0.3044	-

Table 9: The automatic evaluation results of 7 Legal LLMs, GPT-4 and ChatGPT. We use bold to indicate the best and underline to indicate the second-best. Except for Miss, where smaller is better, for other metrics, larger is better.

and ID numbers has been anonymized in the process. Therefore, we can proceed with annotating these two datasets for manual evaluation.

B.2 Rules and Standards of Manual Evaluation

Before starting the annotation process of manual evaluation, we formulated annotation guidelines for the Judicial Reasoning Generation and Legal Consultation tasks through discussions with legal experts.

For the Judicial Reasoning Generation task, the criteria are completeness, relevance and accuracy.

- Completeness: Whether the reasoning content is complete, including the completeness of the reasoning structure and whether explicit penalties are provided.
- Relevance: The degree of relevance between the reasoning content and the case.
- Accuracy: Whether the reasoning content is accurate, including the presence of fabricated facts, incorrect citation of legal provisions, and usage errors.

As for the Legal Consultation task, the criteria include flueny, relevance and comprehensibility.

- Fluency: The fluency and coherence of the response content.
- Relevance: The relevance of the response content to legal issues and its alignment with legal practicality.
- Comprehensibility: The level of understanding of legal issues in the response content.

Additionally, to facilitate computer processing, we standardized the annotation rules for legal experts. For each sample, if the output of the target

Capability	Task	Metrics	Baichuan2-Chat	Baichuan	ChatGLM	Llama-7B	Llama-13B	Llama2-Chat	Chinese-LLaMA-7B	Chinese-LLaMA-13B	Ziya-LLaMA
BIR	Legal Article Recommendation	Acc	0.5620	0.1800	0.7320	0.1750	0.2660	0.4800	0.3790	0.3580	0.6540
		Miss	0.0020	0.5770	0.0030	0.6670	0.2770	0.0170	0.0470	0.0470	0.0020
		F1	0.4507	0.1781	0.7255	0.1953	0.2816	0.4824	0.2439	0.3034	0.6639
	Element Recognition	Acc	0.5400	0.0330	0.4900	0.0370	0.1870	0.1420	0.1310	0.0300	0.5930
		Miss	0	0.6200	0.0110	0.5250	0.0240	0	0.0250	0.9080	0
		F1	0.5218	0.0287	0.4982	0.0143	0.0766	0.1193	0.0745	0.0547	0.5842
		Mcc	0.4995	-0.0629	0.4511	0.0054	-0.0017	0.0872	0.0293	0.0521	0.5427
	Named Entity Recognition	Entity-Acc	0.4731	0	0.0106	0	0	0.0019	0	0	0.4894
	Judicial Summarization	ROUGE-1	0.3584	0.3911	0.5613	0.1655	0.1388	0.2098	0.4094	0.1259	0.5115
		ROUGE-2	0.1632	0.1650	0.2994	0.0584	0.0524	0.1063	0.2174	0.0236	0.2738
		ROUGE-L	0.2785	0.2507	0.4253	0.1180	0.1071	0.1575	0.2963	0.0824	0.3803
	Case Recognition	Acc	0.9700	0.6380	0.8735	0.2235	0.5290	0.8360	0.5235	0.6430	0.9470
		Miss	0.0030	0	0.0940	0.5130	0.0395	0	0.1450	0	0.0010
		F1	0.9714	0.5845	0.9127	0.2323	0.4680	0.8317	0.4897	0.6156	0.9473
LFI	Controversy Focus Mining	Acc	0.0621	0.0556	0.0948	0.0425	0.0588	0.0098	0.0229	0.0621	0.0915
		Miss	0.2941	0.1405	0.7092	0.183	0.2059	0.6863	0.6373	0.1732	0.0327
		F1	0.0412	0.0174	0.1418	0.0131	0.0186	0.0074	0.0202	0.0328	0.0564
		Mcc	0.0186	-0.0061	0.1105	-0.0198	0.0059	-0.0206	-0.0020	0.0069	0.0052
	Similar Case Matching	Acc	0.0154	0	0.5500	0	0	0	0.0038	0.0269	0.0038
		Miss	0.9692	1	0	1	1	1	0.9962	0.9538	0.9962
		F1	0.0299	0	0.3903	0	0	0	0.0076	0.0505	0.0076
	Charge Prediction	Acc	0.2406	0.0060	0.6010	0.4317	0.4643	0.3857	0.3362	0.1391	0.5998
		Miss	0	0.9794	0.2902	0.2273	0.1016	0.2648	0.3277	0.6784	0.0073
		F1	0.1750	0.0118	0.6757	0.3519	0.3679	0.3879	0.3179	0.2021	0.5318
	Prison Term Prediction	Acc	0.7249	0.0745	0.4155	0.0229	0.0458	0.0860	0.0745	0.1003	0.5616
		Miss	0	0	0.0630	0.7393	0.6762	0.1232	0	0	0
		F1	0.6143	0.0103	0.4484	0.0103	0.0580	0.0731	0.0103	0.0533	0.5562
		Mcc	0.0533	0	0.0871	0.0040	0.0096	-0.0347	0	0.0539	-0.0377
	Civil Trial Prediction	Acc	0.6875	0.7037	0.2334	0.4200	0.3063	0.5750	0.7262	0.7113	0.2787
		Miss	0.0013	0.0875	0.6512	0.4537	0.6050	0.1562	0.0525	0.0525	0.0063
		F1	0.6791	0.6450	0.3302	0.4915	0.4046	0.6209	0.6524	0.6446	0.3607
		Mcc	0.1544	0.0196	-0.0403	0.0022	0.0061	0.1081	-0.0064	-0.0275	-0.0348
	Legal Question Answering	Acc	0.3836	0.2304	0.2491	0.1193	0.0772	0.0164	0.1591	0.1497	0.2608
		Miss	0.0152	0.1368	0.0234	0.3519	0.6386	0.9404	0.2070	0.3988	0.0012
		F1	0.3824	0.2432	0.2386	0.0574	0.0557	0.0259	0.0863	0.1660	0.2538
CLA	Judicial Reasoning Generation	ROUGE-1	0.6967	0.5295	0.5096	0.0088	0.1663	0.4052	0.3692	0.2602	0.4113
		ROUGE-2	0.3938	0.2974	0.2158	0.0033	0.0616	0.1759	0.1633	0.1053	0.1948
		ROUGE-L	0.4878	0.3811	0.3363	0.0062	0.1077	0.2816	0.2578	0.2004	0.2975
	Case Understanding	ROUGE-1	0.8249	0.3857	0.8821	0.5995	0.7009	0.7175	0.6745	0.7718	0.8562
		ROUGE-2	0.7920	0.2574	0.8480	0.4948	0.5912	0.6584	0.5441	0.6717	0.8150
		ROUGE-L	0.8219	0.3707	0.8769	0.5880	0.6784	0.7093	0.6507	0.7510	0.8477
	Legal Consultation	ROUGE-1	0.5882	0.2508	0.5007	0.1496	0.1555	0.2618	0.1912	0.1699	0.3494
		ROUGE-2	0.2547	0.0973	0.2022	0.0500	0.0505	0.0885	0.0664	0.0586	0.1529
		ROUGE-L	0.3963	0.2071	0.3478	0.1283	0.1343	0.1793	0.1568	0.1434	0.2554

Table 10: The automatic evaluation results of baseline LLMs.

Model	Judicial Reasoning Generation			Legal Consultation		
	WR_A	WR_B	WR_C	WR_A	WR_B	WR_C
GPT-4	0.34	0.22	0.58	0.98	0.88	0.68
ChatGPT	0.22	0.18	0.66	0.82	0.90	0.66
Fuzi-Mingcha	0.74	0.26	0.94	0.40	0.72	0.40
HanFei	0.58	0.34	0.86	0.34	0.38	0.26
LexiLaw	0.18	0.28	0.48	0.22	0.26	0.24
Lawyer-LLaMA	0.18	0.12	0.60	0.46	0.74	0.32

Table 11: The win rate (WR) of LLMs for the Judicial Reasoning Generation and Legal Consultation tasks. Subscripts A, B, C represent the judgment results of three experts respectively.

LLM is better than the baseline, it is marked as 1; otherwise, it is marked as 0.

During the annotation process, we imported the annotated data into Excel. Each row represents the input for one data point and the outputs of different models. To prevent potential subjective biases from experts toward LLMs, we adopted a model-anonymous annotation approach. Specifically, for each row, we shuffled the order of models, and the shuffling results varied, ensuring that experts wouldn't know which LLM produced the output during annotation.

Finally, we organized the expert annotations to calculate the win rate for each LLM. Figure 2 illustrates the annotation results of expert A for the Judicial Reasoning Generation task.

B.3 Risk Statement of Manual Evaluation

This work is solely intended for academic research and strictly prohibited for any other commercial activities. Before the annotation process, due to the sensitivity of the legal field, we confirmed the usability and security of the dataset and legal experts have conducted ethical evaluations. Additionally, legal experts have conducted ethical evaluations.

B.4 Annotators of Manual Evaluation

The three legal experts conducting the annotations are three graduate students from our research team, specializing in the field of criminal law.

C More Details of Evaluation Metrics

For classification tasks, we select accuracy (Acc), miss rate (Miss), F1 score (F1), and matthews cor-

1	ChatGPT	Fuzi-Mingcha-6B	HanFei-7B	Lawyer-LLaMA-13B	LexiLaw-6B	GPT-4
2	0	1	1	1	1	0
3	1	1	1	0	1	1
4	0	1	0	1	0	1
5	1	1	1	1	1	1
6	0	0	0	0	0	0
7	0	0	1	0	1	0
8	0	1	1	0	1	1
9	0	1	1	0	0	0
10	0	1	1	0	0	0
11	1	1	1	0	0	1
12	0	0	0	1	0	0
13	0	1	1	0	0	1
14	0	1	1	0	0	0
15	1	1	0	0	0	1
16	0	0	0	0	0	0
17	0	1	1	0	0	0
18	0	0	1	0	0	0
19	0	0	0	0	0	0
20	0	1	1	0	0	0
21	0	1	0	0	0	0
22	0	1	1	0	0	0
23	1	0	1	0	0	1
24	0	1	1	0	0	0
25	0	0	0	0	0	0
26	0	1	1	0	0	0
27	0	1	1	0	0	0
28	0	1	1	0	0	0
29	0	0	1	1	1	0
30	0	1	1	0	1	0
31	0	0	0	0	0	0
32	1	1	1	0	0	1
33	0	1	0	1	0	1
34	0	1	0	0	0	1
35	0	1	0	0	0	1
36	1	1	0	1	0	1
37	0	1	1	0	0	0
38	0	1	1	0	0	0
39	1	1	1	0	1	0
40	0	1	1	0	0	0
41	0	1	0	0	0	0
42	0	1	0	0	0	0
43	1	1	0	1	1	1
44	1	0	0	1	0	1
45	0	0	1	0	0	0
46	0	1	0	0	0	0
47	0	0	1	0	0	0
48	1	1	0	0	0	1
49	0	1	0	0	0	0
50	0	1	1	0	0	1
51	0	1	0	0	0	0

Figure 2: The annotation results of expert A for the Judicial Reasoning Generation task. And this annotation is based on using the reference answer as the baseline.

relation coefficient (Mcc) as evaluation metrics for these tasks.

The F1 values presented in our work are all weighted F1.

The miss rate (Miss) is the proportion of missed samples to the total number of test samples. Like MMLU(Hendrycks et al., 2020), we give the candidate categories in the prompt of LLMs for classification tasks. Therefore, for a particular sample, if the outputs of LLMs do not give the results related to the candidate categories, we consider the LLMs have missed that sample, which also means LLMs do not understand the questions.

Finally, as shown in Table 2, the labels of some classification tasks are significantly unbalanced, mirroring real-world scenarios in judicial practice. Relying solely on the F1 score may not effectively reflect the actual performance of LLMs(Chicco and Jurman, 2020). Therefore, we utilize the Matthews correlation coefficient (MCC) to further evaluate the ability of LLMs to handle imbalanced data.

The accuracy of the LLMs in identifying every legal entities (Entity-Acc) is used to evaluate named entity recognition tasks.

For named entity recognition tasks, we use the accuracy of the LLMs in identifying every legal entities (Entity-Acc).

For text generation tasks, we use ROUGE as evaluation metrics for this task, since ROUGE remains one of the mainstream evaluation metrics for LLMs(Fei et al., 2023; Srivastava et al., 2022).

D More Results

Model	JRG_{ref}	LC_{ref}
GPT-4	0.57	0.77
ChatGPT	0.55	0.69
Fuzi-Mingcha	0.52	0.59
HanFei	0.55	0.71
LexiLaw	**0.63**	**0.80**
Lawyer-LLaMA	0.53	0.52

Table 12: The agreement scores of LLMs. JRG and LC represent the Judicial Reasoning Generation and Legal Consultation tasks, respectively. The subscript $_{ref}$ indicates the agreement of the evaluations from the three experts when using the reference answer as the baseline.

D.1 The Automatic Evaluation Results

As shown in Table 9 and Table 10, we can observe that their performance is consistent with the trend of our score results. GPT-4 and ChatGPT have strong multi-level capabilities, with a certain legal logic, while other LLMs have strong text generation capabilities but lack legal logic.

These detailed tables can also help us more clearly identify the strengths and weaknesses of LLMs in various tasks. The legal LLMs performed unsatisfactorily in tasks corresponding to the major and minor premises in syllogism, such as Legal Article Recommendation and Element Recognition. They also fell short in further reasoning tasks such as Charge Prediction, Prison Term Prediction, and Civil Trial Prediction compared to GPT-4 and ChatGPT. Overall, the performance of these LLMs indicates a lack of information retrieval and reasoning related to legal logic.

Task	Evaluation	GPT-4	ChatGPT	Fuzi-Mingcha	HanFei	LexiLaw	Lawyer-LLaMA	τ	p
Judicial Reasoning Generation	Automatic	3	4	2	1	6	5	0.7333	0.0566
	Manual	3	4	1	2	5	6		
Legal Consultation	Automatic	2	1	3	5	6	4	0.8281	0.0217
	Manual	1	2	3	5	6	3		

Table 13: The agreement scores for manual and automatic evaluation.

D.2 The Win Rate of LLMs for Each Expert

As shown in Table 11, Expert A and B have similar win rates, while Expert C differs significantly from them. This suggests that while legal logic is commonly recognized among legal experts, there are still individual differences in actual judgment, influenced by certain subjectivity.

D.3 The Agreement Scores for Expert Evaluation

Furthermore, for the manual evaluation, we calculated agreement scores for expert evaluation, as shown in Table 12. Based on this, we observe the following fact:

Although experts can find the lack of legal logic in LLMs, assessing legal logic may also pose a challenge for experts. The agreement score for the Judicial Reasoning Generation task is noticeably lower than that for the Legal Consultation task. The reference answers for judicial reasoning generation tasks are derived from actual court judgments in legal documents, serving as the gold answers. This task emphasizes the completeness and accuracy of formal content, which is directly related to legal logic. This allows experts to judge based on their legal logic, which may be affected by their legal background, bring noise, and also bring challenges to evaluation.

On the other hand, legal consultation work involves legal opinions for the public, covering a broader range of legal areas but addressing common legal issues. Experts provide answers more based on fluency rather than based on the legal logic of legal practice. This makes it easier for experts to judge, and the agreement scores are higher.

D.4 The Agreement Scores for Manual and Automatic Evaluation

We ranked the LLMs evaluated automatically based on the scores in Table 6, and ranked the LLMs evaluated manually based on the average win rate scores in Table 7. Subsequently, we calculated Kendall's tau scores (τ) and significance values (p) for both Judicial Reasoning Generation and Legal Consultation tasks, as shown in Table 13. We observe that for these same LLMs, two entirely different evaluation methods demonstrate similar rankings, both with high τ values. Thus, this further strengthens the reliability of our automatic evaluation and confirms the conclusions summarized in section 5.3.