LAiW: A Chinese Legal Large Language Models Benchmark

Yongfu Dai[a], **Duanyu Feng**[a], **Jimin Huang**[b], **Haochen Jia**[a], **Qianqian Xie**[c], **Yifang Zhang**[a], **Weiguang Han**[c], **Wei Tian**[d], **Hao Wang**[a,*]
[a]Sichuan University, Chengdu, China. [b]The Fin AI, Singapore.
[c]Wuhan University, Wuhan, China.[d]Southwest Petroleum University, Chengdu, China.
wal.daishen@gmail.com fengduanyuscu@stu.scu.edu.cn jimin.huang@thefin.ai
wwx990211@gmail.com qianqian.xie@thefin.ai zhangyf_ivy@foxmail.com
han.wei.guang@whu.edu.cn 418818347@qq.com wangh@scu.edu.cn

Abstract

General and legal domain LLMs have demonstrated strong performance in various tasks of LegalAI. However, their current evaluations lack alignment with the fundamental logic of legal reasoning, the legal syllogism. This hinders trust and understanding from legal experts. To bridge this gap, we introduce LAiW [1], the Chinese legal LLM benchmark structured around the legal syllogism. We evaluate legal LLMs across three levels of capability, each reflecting a progressively more complex stage of legal syllogism: fundamental information retrieval, legal principles inference, and advanced legal applications, and encompassing a wide range of tasks in different legal scenarios. Our automatic evaluation reveals that LLMs, despite their ability to answer complex legal questions, lack the inherent logical processes of the legal syllogism. This limitation poses a barrier to acceptance by legal professionals. Furthermore, manual evaluation with legal experts confirms this issue and highlights the importance of pretraining on legal text to enhance the legal syllogism of LLMs. Future research may prioritize addressing this gap to unlock the full potential of LLMs in legal applications.

1 Introduction

With the emergence of ChatGPT and GPT-4 and their excellent text processing capabilities (Zhao et al., 2023), researchers begin to pay considerable 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 studies how artificial intelligence can assist in legal practice (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 the LLMs, legal experts are cautious about their practical application in law (Dahl et al., 2024). They believe that the lack of logical process, the **legal syllogism**, of LLMs in legal practice may significantly affect the fairness in legal practice[2]. Nevertheless, 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 pretraining (Wen and He, 2023) or fine-tuning (Cui et al., 2023a) to acquire knowledge or capabilities in the legal field, most of them prioritize improving LLM performance in these tasks, neglecting the underlying logic of the legal syllogism. Only SLJA (Deng et al., 2023) offers a legal syllogism dataset for continue pre-training LLM Fuzi-Mingcha (Wu et al., 2023), but it is limited in the tasks of legal judgments. Existing benchmarks for evaluating these models are also constructed based on the performance of the models in individual tasks such as legal question and answer and consultation (Yue et al., 2023; Fei et al., 2023; Zhong et al., 2020b; Choi, 2023; Steenhuis et al., 2023). This fails to fully reflect the applications and the legal syllogism of the LLMs by legal practitioners. Therefore, it is important to explore the abilities of the LLMs from the perspective of the legal syllogism in law to ensure that legal practitioners can better understand and use the LLMs.

More precisely, the **legal syllogism** is the core legal reasoning ability recognized by legal experts, involving obtaining evidence and legal articles, making conclusions, and their interconnections (Kuppa et al., 2023; Trozze et al., 2023), as shown in Table 1. First, the ability to extract information from the legal texts, then the ability to provide a reliable

*This is the corresponding author.
[1]https://github.com/Dai-shen/LAiW
[2]https://github.com/liuchengyuan123/LegalLLMEvaluation/

Proceedings of the 31st International Conference on Computational Linguistics, pages 10738–10766
January 19–24, 2025.

Stage	Explanation	Example
Major Premise	Legal norms	The intentional murderer should be sentenced to death.
Minor Premise	Case facts	A intentionally killed B.
Conclusion	Legal judgment	A should be sentenced to death.

Table 1: The definitions and examples of legal syllogism, illustrated with a clear and straightforward example. Legal syllogism is a step-by-step logical reasoning process, organized into multiple levels of complexity.

and reasoned answer based on solid legal knowledge, and ultimately the ability to form a complete response. This entire process avoids logical confusion and ensures the preservation of the legal logic and the reliability of the conclusions.

In this work, to investigate the **legal syllogism** of LLM, we propose the Chinese legal LLM benchmark LAiW[3]. We categorize the legal capabilities of LLMs into three levels: fundamental information retrieval (FIR), legal principles inference (LPI), and advanced legal applications (ALP), each reflecting a progressively more complex stage of legal syllogism. In the FIR stage, we assess whether the LLMs can extract legal provisions and legal evidence from the given legal text, corresponding to obtaining the minor premise and major premise of the legal syllogism. Then, in the LPI stage, we verify if the LLMs can derive a preliminary conclusion based on these premises identified in the previous stage, corresponding to making a conclusion of the legal syllogism. Finally, the ALP stage of our benchmark examines how LLMs apply the legal syllogism in real-world legal practice. This involves analyzing specific case facts within the context of legal norms and drawing conclusions based on this application. To capture these capabilities, we curated 14 tasks from existing LegalAI tasks, reconstructing them to reflect this complex reasoning process.

The analysis of our benchmark includes both automatic and manual evaluations to assess LLMs. While automatic evaluation reveals strong text generation skills in advanced legal applications, it exposes a lack of logical rigor in fundamental information retrieval and legal principles inference. Manual evaluations by legal experts confirm this, highlighting the discrepancy between apparent legal reasoning and actual adherence to the legal syllogism. This suggests a need for pre-training to instill the syllogistic logic in LLMs, as fine-tuning alone is insufficient. This insight may guide future improvements for LLMs in the legal domain.

[3]It means "AI in LAW".

Our contributions are as follows:

- We are proud to introduce the Chinese legal LLMs benchmark LAiW, which is designed based on the legal syllogism. We categorize the legal capabilities of the LLMs into three levels to facilitate a more precise evaluation of the LLMs in legal practice and to enhance legal experts' understanding of the LLMs.
- Based on our automatic evaluation, we demonstrate that current legal LLMs do not have legal syllogism. Though the LLMs demonstrate strong text generation abilities to advanced legal application, they struggle 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 the LLMs. The results indicates the need of pretraining on legal text with the legal syllogism for the LLMs for future improvement.

2 Related Work

Chinese Legal LLMs. Table 2 summarizes current Chinese legal LLMs and some general models. Many of these LLMs prioritize practical legal applications, fine-tuned on legal datasets. Examples include LawGPT_zh (Liu et al., 2023), Lawyer-LLaMA (Huang et al., 2023a), ChatLaw (Cui et al., 2023a), and LexiLaw, which excel in answering legal questions and providing consultations. However, they often rely on external knowledge bases to compensate for their limited legal knowledge, potentially impacting accuracy and comprehensiveness. Other LLMs, like LaWGPT (Pengxiao et al., 2023), wisdomInterrogatory, Fuzi-Mingcha (Wu et al., 2023; Deng et al., 2023), and HanFei (Wen and He, 2023), employ pre-training or continue pre-training to enhance their legal understanding, covering a wider range of tasks like element extraction and case classification. While these advancements improve overall effectiveness in legal applications, a critical shortcoming remains: only Fuzi-Mingcha use legal syllogisms dataset (with limited scope encompassing a few aspects of legal judgment analysis) for continue pre-training, and many LLMs may largely overlook the essential logical framework of the legal syllogism, which is of paramount importance to legal professionals.

Model	Model Size	Model Domain	From	Baseline	Creator	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 2: The LLMs evaluated in our work. LaWGPT and wisdomInterrogatory undergo pretraining 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.

Benchmark	Tasks	Size	Taxonomy	Legal Perspective	Expert Evaluation
DISC-Law-Eval	7	3k	Task Difficulty	×	×
LawBench	20	10k	Bloom's Cognitive Model	×	×
LAiW	**16**	**11k**	**Legal Syllogism**	√	√

Table 3: Comparison of different Chinese legal benchmarks. "Legal Perspective" refers to whether the construction of benchmarks are mainly guided from legal perspective, and "Expert Evaluation" refers to whether legal experts manually evaluate the LLMs with the benchmark.

Legal LLMs Benchmark. LegalAI has spurred the development of numerous tasks combining law and computer science, from NLP-focused tasks like legal NER and summarization (Kanapala et al., 2019) to legal-focused tasks like similar case matching (Locke and Zuccon, 2022; Sansone and Sperlí, 2022). From a legal perspective, LegalAI also encompasses the legal syllogism, from legal element 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). These tasks provide ample data for evaluating Chinese legal LLMs (Zhong et al., 2020b). While recent benchmarks have acknowledged the importance of legal syllogisms in legal reasoning (Yue et al., 2023), they still face limitations. Benchmarks like LawBench (Fei et al., 2023) and DISC-Law-Eval (Yue et al., 2023) are still constructed from an AI perspective, focusing on evaluating LLMs' knowledge-based abilities (as shown in Table 3).All these benchmarks also lack manual evaluation, which impedes the identification of potential improvements and a deeper understanding of LLMs' capabilities. Furthermore, existing non-Chinese legal benchmarks, like LexGLUE (Chalkidis et al., 2023), LEXTREME (Niklaus et al., 2023), and LegalBench (Guha et al., 2023), align with the common law system, emphasizing case law. This contrasts with the civil law system, which relies on statutory provisions and necessitates a grounding in the legal syllogism. Our work addresses this gap by focusing on evaluating LLMs through the lens of the legal syllogism, specifically within the whole Chinese civil law system.

3 Benchmark Construction

This section categorizes the abilities of LLMs for legal tasks using the practical application of the legal syllogism. We then introduce our benchmark, LAiW, for evaluating Chinese legal LLMs, struc-

Capability	Task	ID	Primary Origin Dataset	LAiW	Domain	Task Type	Class	Balance
FIR	Legal Article Recommendation	F1	CAIL2018 (Xiao et al., 2018)	1000	Criminal	Classification	3	0.231
	Element Recognition	F2	CAIL-2019 (Zhang et al., 2022a)	1000	Civil	Classification	20	0.002
	Named Entity Recognition	F3	CAIL-2021 (Cao et al., 2022)	1040	Criminal	Named Entity Recognition	-	-
	Judicial Summarization	F4	CAIL-2020 (Huang et al., 2023b)	364	Civil	Text Generation	-	-
	Case Recognition	F5	CJRC (Duan et al., 2019)	2000	Criminal, Civil	Classification	2	0.499
LPI	Controversy Focus Mining	L1	LAIC-2021	306	-	Classification	10	0.029
	Similar Case Matching	L2	CAIL-2019 (Xiao et al., 2019)	260	Civil	Classification	2	0.450
	Charge Prediction	L3	Criminal-S (Hu et al., 2018)	827	Criminal	Classification	3	0.172
	Prison Term Prediction	L4	MLMN (Ge et al., 2021)	349	Criminal	Classification	3	0.074
	Civil Trial Prediction	L5	MSJudeg (Ma et al., 2021)	800	Civil	Classification	3	0.065
	Legal Question Answering	L6	JEC-QA (Zhong et al., 2020c)	855	-	Classification	4	0.201
ALA	Judicial Reasoning Generation	A1	AC-NLG (Wu et al., 2020)	834	Civil	Text Generation	-	-
	Case Understanding	A2	CJRC (Duan et al., 2019)	1054	Criminal, Civil	Text Generation	-	-
	Legal Consultation	A3	CrimeKgAssitant (Liu et al., 2023)	916	-	Text Generation	-	-

Table 4: Statistical information of our dataset. All datasets are sourced from open-source. In the 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.

tured around these three ability levels. To ensure a thorough assessment, we employ both automated evaluation with quantifiable metrics and manual evaluation by legal professionals.

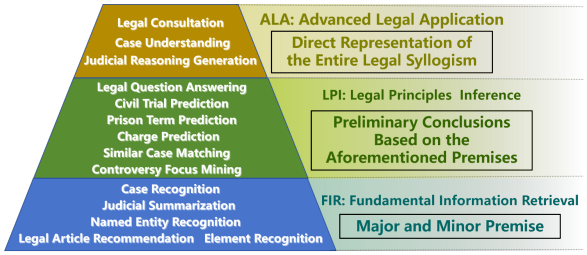


Figure 1: Multi-level Legal Capabilities of LLMs.

3.1 The Legal Syllogism for LLMs

The syllogism, traditionally used to analyze criminal cases in civil law systems (Wróblewski, 1974; Patterson, 2013), has expanded its reach in recent years to encompass civil cases as well (Wang, 2022). This has led to its widespread adoption in nearly all areas of Chinese law, establishing it as a universal framework for legal reasoning.

This framework involves three components: the major premise, which represents the applicable legal articles; the minor premise, which establishes the specific facts of the case through evidence analysis; and the conclusion, which forms the legal judgment based on the preceding premises. As illustrated in Table 1, legal practice essentially involves connecting legal articles (major premise) to the unique circumstances of each case (minor premise) to arrive at a legal decision (conclusion). This interconnectedness highlights the intricate interplay between legal articles and specific facts.

To ensure that LLMs operate within a logical framework consistent with legal practice, we categorize their capabilities into three levels, aligning them with the legal syllogism as shown in Figure 1. By combining the skills of acquiring both minor and major premises, we establish the "**fundamental information retrieval**" level. Building on this foundation, we develop the "**legal principles inference**" level, enabling LLMs to draw preliminary conclusions based on these acquired premises. Finally, to evaluate the full process of the legal syllogism, we introduce the "**advanced legal application**" level. This level assesses the LLMs' ability to apply the full syllogistic framework to complex legal problems.

3.1.1 FIR: Fundamental Information Retrieval

The Fundamental Information Retrieval level comprises five tasks[4] designed to evaluate LLMs' foundational abilities in processing legal text. These tasks focus on identifying key elements related to both minor and major premises, such as legal evidence, knowledge, and categorization. They serve as the initial step in the legal syllogism framework, laying the groundwork for subsequent reasoning by gathering necessary elements.

Therefore, this level includes Legal Articles Recommendation, which identifies relevant legal arti-

[4]For detailed task selection criteria and definitions of each task, please refer to Appendix A.1 and A.2, respectively.

cles (major premises), and Elements Recognition, which pinpoints crucial elements (minor premises) from case facts. Additionally, three established NLP tasks are included: Named Entity Recognition, Judicial Summarization, and Case Recognition, which extract key information and classify cases. While these tasks don't require extensive legal knowledge, they provide valuable text-based insights that are essential for both legal and computational applications.

3.1.2 LPI: Legal Principles Inference

The Legal Principles Inference level evaluates LLMs' ability to apply legal reasoning, bridging the gap between minor and major premises to draw basic conclusions and judgments. This level is crucial to the legal syllogism, connecting its component parts.

We structure this level into three categories with six tasks: (1) Basic Legal Applications. Controversial Focus Mining, identifies key points of contention in civil law cases based on facts and legal articles; Similar Case Matching, finds similar cases as references to ensure fairness in judgment. (2) Predicting Legal Outcomes. Charge Prediction (Criminal Law), predicts charges based on criminal cases; Prison Term Prediction (Criminal Law), predicts potential sentences in criminal cases; Civil Trial Prediction (Civil Law), predicts outcomes in civil cases. (3) Legal Question Answering. Requires LLMs to integrate legal knowledge and provide basic legal responses based on given facts.

These tasks assess LLMs' ability to synthesize information and make basic legal inferences, demonstrating their understanding of legal articles and their application to specific case scenarios.

3.1.3 ALA: Advanced Legal Application

The Advanced Legal Application level probes the depths of complex legal reasoning, investigating whether LLMs can effectively utilize the entire process of legal syllogism to tackle challenging tasks. This level aims to determine if LLMs can go beyond obtaining legal premises and drawing conclusions separately, simulating the whole process of legal professionals. To achieve this, we propose three challenging tasks, Judicial Reasoning Generation, Case Understanding, and Legal Consultation, requiring LLMs to demonstrate their grasp of the legal syllogism inherent in legal judgments.

Judicial Reasoning Generation requires LLMs to recreate the full logical process of legal judgments, from premises to conclusions. Case Understanding focuses on comprehending the logic behind legal cases. Legal Consultation involves using this understanding to provide advice like a legal professional.

3.2 Datasets Construction

With the mentioned criteria for the division of capabilities and tasks, we construct the evaluation dataset for our LAiW benchmark based on the open-source datasets. This dataset is divided into two parts: Automatic and Manual, reflecting the different methods used for evaluation.

3.2.1 Automatic Evaluation Datasets

We've developed datasets for all 14 tasks[5] that can be automatically assessed, shown in Table 4. These datasets are primarily drawn from the CAIL competition data (Xiao et al., 2018; Zhang et al., 2022a; Huang et al., 2023b) and commonly used open-source data (Ge et al., 2021; Wu et al., 2020; Liu et al., 2023). We've included a diverse range of legal areas, encompassing criminal, civil, constitutional, social, and economic law, to cover a broad spectrum of legal scenarios.

To ensure LLMs can provide relevant answers, we designed specific prompts for each task. These prompts were carefully crafted, using ChatGPT to ensure their quality, and validated by legal experts to confirm their accuracy.

3.2.2 Manual Evaluation Datasets

Our automatic evaluation results (Section 5.2) indicate that the LLMs we evaluated struggle to adhere to the principles of legal syllogism. While LLMs appear to possess advanced legal application capabilities, their performance in following the structured framework of legal syllogism falls short. To delve deeper into this observation and understand the underlying reasons, we conducted a manual evaluation specifically focusing on the third level (Advanced Legal Application).

Given the cost of manual evaluation, we focused on two tasks most closely tied to legal syllogism: Judicial Reasoning Generation and Legal Consultation. These tasks represent the application of legal syllogism for legal professionals and the general public, respectively.

[5]Examples and the detailed processing methods can be found in Appendix B.

4 Evaluation for Benchmark

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

4.1 Automatic Evaluation

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

Table 5: The metrics for automatic evaluation.

The automatic evaluation includes the tasks of classification, named entity recognition and text generation . Table 5 presents the evaluation metrics[6] for each task.

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

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

The total score is computed by averaging the scores of the three levels of capabilities. These level scores, in turn, are determined by averaging the task scores within each level.

4.2 Manual Evaluation

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

Table 6: The assessment criteria for manual evaluation.

To ensure reliable assessment, we discussed the criteria[7] with the legal experts who conduct the manual evaluation. We adopted the approach used in (Dubois et al., 2023; Li et al., 2023) for manual evaluation, shown in Table 6. Such approach considers legal experts as evaluators and use reference answers as the baseline to compute the win rate for the target LLMs. For example, when using the reference answer as the baseline, legal experts assess the output of the target LLM and the reference answer from multiple dimensions of judgment, and then choose the most satisfactory response.

5 Experiment

In this section, we present the experiment settings and highlight the key results of the legal syllogism in the 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) and 6 corresponding baseline LLMs (Du et al., 2022; Cui et al., 2023c; Zhang et al., 2022a), and 5 effective general LLMs (Baichuan, 2023; Touvron et al., 2023a) such as GPT-4 (gpt-4-1106-preview) and ChatGPT (gpt-3.5-turbo-16k-0613). For a fair evaluation, all LLMs were evaluated without the addition of RAG (Retrieval-Augmented Generation) modules. Table 2 lists more detailed information about these LLMs. We use the greedy generation strategy across all of these LLMs to ensure reproducibility of results.

For the manual evaluation, we choose the four top-performing legal LLMs. They are FuzMingcha (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

Table 7 presents the scores for each level and the overall score of our automatic evaluation [8]. We analyze these results from three perspectives: overall performance, the legal syllogism of Chinese Legal LLMs, and an exploration of in-context learning's impact on the legal syllogism of LLMs.

Overall results. Our evaluation reveals a significant gap between current open-source LLMs and specifically trained legal LLMs, particularly when compared to GPT-4 and ChatGPT. Table 7 shows that GPT-4 and ChatGPT consistently outperform all other models, achieving top scores across most tasks. This superiority extends to various levels of evaluation, indicating a clear advantage in their overall capabilities. Among the

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

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

[8]More detailed results for each task are provided in Appendix E.1. We also evaluated several pre-trained language models (PLMs); however, due to space limitations and the focus of this study, the detailed results are presented in Appendix E.5.

Model	Fundamental Information Retrieval						Legal Principles Inference							Advanced Legal Application				Total Score
	F_1	F_2	F_3	F_4	F_5	Avg.	L_1	L_2	L_3	L_4	L_5	L_6	Avg.	A_1	A_2	A_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 7: Scores for LLMs across various levels of the LAiW benchmark (calculated using Equation 1). The top five overall performing LLMs are highlighted in bold. Task names for each level are detailed in Table 4.

open-source LLMs, only Baichuan2-Chat, ChatGLM, and Ziya-LLaMA attain a total score of 45 or higher. However, their performance in the FIR and LPI levels (basic legal logic and knowledge) lags significantly behind GPT-4 and ChatGPT. For the top four specifically trained legal LLMs (Fuzi-Mingcha, HanFei, LexiLaw, and Lawyer-LLaMA), are even lower than 45.

These discrepancies are likely due to two primary factors: (1) *Model Size*: GPT-4 and ChatGPT boast significantly larger parameter counts, providing them with greater capacity for learning and generalization. (2) *Data Exposure*: GPT-4 and ChatGPT may have been trained on a broader dataset during pretraining, including a wider range of legal data across multiple languages. In contrast, the open-source LLMs we selected primarily target the Chinese community, potentially limiting their exposure to diverse legal information.

The Legal Syllogism of Chinese Legal LLMs. Our analysis reveals a significant gap between the capabilities of Chinese Legal LLMs and the legal syllogism. While these models excel at advanced legal applications, they struggle with tasks in other basic levels. Table 7 highlights this discrepancy, showing that most legal LLMs score nearly 20 points higher in the ALA level (direct logic application) compared to the FIR and LPI levels (basic legal logic and knowledge).

This stark contrast contradicts the typical logical structure of law. It suggests that these LLMs have primarily learned to generate legal texts without truly grasping the underlying legal logic. Consequently, they struggle to identify the major and minor premises needed for legal syllogism, limiting their ability to reach sound conclusions.

However, ChatLaw stands out among the legal LLMs, demonstrating strong performance in the FIR level. This likely stems from the robust performance of its base model, Ziya-LLaMA.

Model	FIR	LPI	ALA	Total Score
Baichuan2-13B-Chat	$56.39_{+2.72}$	$45.34_{+13.31}$	$56.08_{-2.32}$	$52.60_{+4.65}$
Baichuan-7B	$30.35_{+9.15}$	$18.13_{+2.67}$	$49.88_{+19.02}$	$32.79_{+10.28}$
ChatGLM-6B	$22.67_{-28.84}$	$12.33_{-24.75}$	$50.85_{-1.59}$	$28.62_{-18.39}$
Llama-7B	$21.17_{+10.05}$	$21.03_{+5.63}$	$34.11_{+11.57}$	$25.44_{+9.08}$
Llama-13B	$18.92_{+1.41}$	$26.04_{+10.96}$	$34.06_{+4.66}$	$26.34_{+5.68}$
Llama2-7B-Chat	$34.49_{+2.53}$	$28.04_{+15.27}$	$43.61_{+4.97}$	$35.38_{+7.59}$
Chinese-LLaMA-7B	$23.60_{+1.28}$	$6.55_{-11.70}$	$37.86_{+3.70}$	$22.67_{-2.24}$
Chinese-LLaMA-13B	$37.18_{+16.16}$	$22.59_{+3.43}$	$40.97_{+6.17}$	$33.58_{+8.59}$
Ziya-LLaMA-13B	$48.40_{-13.07}$	$30.49_{+1.05}$	$46.23_{-0.22}$	$41.71_{-4.08}$
HanFei-7B	$35.86_{-1.56}$	$28.87_{+12.4}$	$47.70_{-5.81}$	$37.48_{+1.72}$
Wisdom-Interrogatory-7B	$36.63_{+23.97}$	$25.83_{+15.38}$	$53.05_{+19.68}$	$38.50_{+19.68}$
Fuzi-Mingcha-6B	$22.67_{-17.01}$	$12.33_{-15.13}$	$50.85_{-3.86}$	$28.62_{-12.00}$
LexiLaw-6B	$30.97_{-10.35}$	$9.56_{+0.68}$	$39.68_{-4.05}$	$26.74_{-4.57}$
LaWGPT-7B	$21.55_{+6.08}$	$12.28_{-1.99}$	$44.63_{+6.31}$	$26.15_{+3.47}$
Lawyer-LLaMA-13B	$49.27_{+18.42}$	$32.49_{+26.10}$	$48.97_{-1.53}$	$43.57_{+14.33}$
ChatLaw-13B	$47.94_{-10.08}$	$34.83_{+22.09}$	$37.94_{+31.20}$	$40.24_{+14.47}$

Table 8: Scores for LLMs across various levels of the LAiW benchmark with in-context learning. Example prompts and answers were provided to guide the LLMs.

The In-context learning for the legal syllogism of LLMs. In-context learning does not consistently improve the capability of legal syllogism for LLMs[9]. In Table 8, while LLMs like Wisdom-Interrogatory-7B indicate a score increase of nearly 20 points through in-context learning, Fuzi-Mingcha-6B and LexiLaw-6B experience decreases of about 12 and 5 points, respectively. This suggests that in-context learning may enhance legal syllogism abilities for certain LLMs but can also interfere with their performance in others. This observation might be linked to the version of the LLMs, suggesting that earlier models possess weaker capa-

[9]More details can be found in Appendix E.6.

bility of the in-context learning. The findings highlight that LLMs may not acquire legal syllogism skills solely through examples in the In-context learning.

Overall, this performance gap shown in this section raises concerns about the current state of Chinese Legal LLMs and their ability to meet the expectations of legal professionals. The weak connection to the legal syllogism framework, a cornerstone of legal reasoning, could undermine trust in these models for legal applications.

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	<u>43.97</u>	**0.85**	0.15
ChatGPT	41.74	0.35	0.27	**48.79**	<u>0.79</u>	0.12
Fuzi-Mingcha	**63.58**	**0.65**	0.35	35.22	0.51	0.19
HanFei	<u>60.13</u>	<u>0.59</u>	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 9: The average win rate (WR) of the LLMs for the tasks of Judicial Reasoning Generation and Legal Consultation. The total score represents the score obtained by the LLMs through automatic evaluation on our benchmark. We use bold to indicate the best and underline to indicate the second-best.

According to the expert evaluation criteria in Section 4.2, Table 9 presents the average win rates from three legal experts.[10] Based on these results, we present three key findings.

Manual evaluation and automatic evaluation share similarities. This enhances the reliability of our automatic evaluation. From Table 9, we observe that the results of the manual evaluation and the automatic evaluation are similar. For instance, in both evaluation rounds, Fuzi-Mingcha (63.58 in automatic evaluation, 0.65 in win rate) and HanFei (60.13 in automatic evaluation, 0.59 in win rate) perform best in the Judicial Reasoning Generation task, while GPT-4 and ChatGPT excel in the Legal Consultation task. This indicates that our automatic evaluation can provide a reliable path for the legal syllogism assessment of the legal LLMs and further reduce the manual effort. Therefore, our assessment of legal syllogism is granular, and the degrees of emphasis on legal syllogism in different scenarios may also be reflected by our automatic evaluation of different tasks.

Additionally, while we employed ROUGE for text generation of automatic evaluation, recognizing its limitations as a metric, our manual evaluation reveal that the ROUGE still demonstrate a degree of competence in reflecting legal syllogism.

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

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

The future of Chinese Legal LLMs. Fine-tuned legal LLMs can improve the normalization of the legal text generation, but they may sacrifice the legal syllogism. Furthermore, for the legal LLMs, undergoing additional pretraining on legal text could be a pathway to acquiring diverse legal capabilities and understanding legal syllogism.

From manual evaluation, legal experts find that the fine-tuned legal LLMs such as Lawyer-LLaMA has the ability of generating texts with good normalization, but may be not good at legal syllogism. Referring to Table 7, we can further find that the acquisition of such ability may stem from the fine-tuning LLMs on ALA-level tasks compared with their base models. This enables the LLMs to respond in a certain standard style, but without the framework of the legal syllogism, such fine-tuning may result in a decline in performance at the FIR and LPI levels. Furthermore, our automated evaluation results also demonstrate that incorporating

[10] Detailed win rate scores for each expert are in Appendix E.2. Appendix E.3 details the agreement scores (consistency) between automatic scores. Appendix E.4 presents agreement scores between automatic and manual scores.

in-context learning may not enhance the capability of legal syllogism for fine-tuned legal LLMs. This reinforces the observation that legal syllogism is not implicitly acquired during the fine-tuning.

On the other hand, the legal LLMs like HanFei and Fuzi-Mingcha, which rely more on pretraining, may indicate how Chinese Legal LLMs acquire ability of the legal syllogism. HanFei and Fuzi-Mingcha, although it is based on an older LLM structure (Bloomz, ChatGLM) with extensive pretraining on legal texts, demonstrates the capabilities on par with subsequent legal LLMs in automatic and manual evaluations. Furthermore, GPT-4 and ChatGPT, which are the models with extensive pretraining on large corpora, also show excellent performance at the FIR and LPI levels. These findings indicate that developing legal reasoning and comprehensive capabilities with like legal syllogism may require pretraining, rather than just fine-tuning.

6 Conclusion

This study introduces a new benchmark for evaluating Chinese Legal LLMs based on the legal syllogism. To match the process of the legal syllogism step by step, the benchmark categorizes the legal capabilities of the LLMs into three levels which encompass a total number of 14 tasks. Both automatic and manual evaluations are conducted in the benchmark evaluations. The results by the automatic evaluations show that existing LLMs excel in text generation for advanced legal application but struggle with basic fundamental information retrieval and legal principles inference, leading to a lack of legal syllogism and distrust among legal experts. Manual evaluations reveal that while the LLMs may bridge the gap in the legal syllogism in some application, they still exhibit significant discrepancies compared with legal experts. This demonstrates the importance and necessity for further pretraining of the LLMs in the legal domain to gain the legal syllogism rather than solely relying on fine-tuning.

Limitations

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

1) In the manual evaluation experiment, to save the workload, only a portion of the data and the LLMs are sampled and chosen for evaluation. This should involve more collaboration 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. This need to develop additional tasks to refine the logic of legal practice at each stage.

3) We acknowledge that prompts might introduce sensitivity for different LLMs, and we have made efforts to reduce their impact in our benchmark. As mentioned in section 3.2, we have strived to ensure that all the LLMs can provide relevant responses to our prompts to guarantee fairness.

Therefore, we also strongly encourage researchers and industry professionals to participate in the development of this benchmark by contributing more tasks and evaluation methods, thus enriching the evaluation of legal syllogistic reasoning.

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

We would like to thank the editors and reviewers for their insightful comments and guidance, which significantly improved this work. This research is supported by the National Key R&D Program of China (No. 2022YFC3301503).

References

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

Sheng Bi, Zafar Ali, Tianxing Wu, and Guilin Qi. 2024. Knowledge-enhanced model with dual-graph interaction for confusing legal charge prediction. *Expert Systems with Applications*, 249:123626.

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.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

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*.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009, Singapore. Association for Computational Linguistics.

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*.

Duanyu Feng, Bing Hu, Yifang Zhang, Wei Tian, and Hao Wang. 2023. Multi-scale heterogeneous graph attention network for prison term prediction. In *2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, volume 3, pages 1395–1404. IEEE.

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.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

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.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

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*.

Zhu Wang. 2022. 司法人工智能推理辅助的"准三段论"实现路径. 政法论坛, 40(05):28–39.

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*.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, and Yanqing An. 2024. A circumstance-aware neural framework for explainable legal judgment prediction. *IEEE Transactions on Knowledge and Data Engineering*.

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. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708.

A More Details of Task Construction

A.1 Construction Criteria

The design of our benchmark, LAiW, is guided by several key principles:

- Alignment with the Legal Syllogism: Tasks directly related to the legal syllogism are assigned to their corresponding capability levels. For example, Legal Article Recommendation aligns with the major premise, Element Recognition with the minor premise, and Charge Prediction with the conclusion.
- Broad Coverage of Legal Domains: The benchmark aims to encompass a wide range of legal scenarios, including both civil and criminal law. This is reflected in tasks like Civil Trial Prediction and Prison Term Prediction, which encompass inferences within both legal domains.
- Diverse User Perspectives: We strive to cater to the needs of various users within the legal system. This includes tasks like Judicial Reasoning Generation, designed for legal professionals, and Legal Consultation, catered towards the general public.
- Inclusion of Open-Source Tasks: Tasks that may be indirectly related to the legal syllogism and are publicly available are also included. This ensures a comprehensive evaluation, incorporating tasks like Named Entity Recognition and Similar Case Matching.

By incorporating these criteria, LAiW provides a multifaceted and robust benchmark for evaluating the legal capabilities of Chinese LLMs. More detailed definitions of the tasks in LAiW are shown in next section.

A.2 Task Definition

In this section, we provide the definitions for the 14 tasks included in our benchmark.

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

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

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

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

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

Controversy Focus Mining: 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.

Similar Case Matching: 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.

Charge Prediction: 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.

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

Civil Trial Prediction: 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.

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

Judicial Reasoning Generation: 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.

Case Understanding: 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.

Legal Consultation: 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.

B More Details of Instruction Dataset

B.1 Data Source

For the convenience of researchers, Table 10 lists the original sources of our reconstructed dataset.

B.2 Data Field

Each instruction dataset is converted to JSONL format. The dataset comprises the following field:

{

id: [integer] a unique identifier for each data sample

query: [string] the input question and prompt

text: [string] the input text content

answer: [string] the expected answer or response

}

Additionally, for the instruction datasets of classification tasks shown in Table 4, we also provide two additional field:

{

choices: [list] a list of possible answer choices

gold: [integer] the correct or gold standard answer

}

B.3 Data Cleaning

First, we constructed templates of LLM queries for each task based on the task definitions in Section A.2. To ensure that most LLMs could correctly respond to our query templates, we created prompts for each task by incorporating feedback from ChatGPT and legal experts. Then, we filtered out low-quality data from the original dataset (including redundant text, disorganized language, excessive symbols, etc.) and designed regular expressions to extract the necessary information for the query templates from the original dataset. We have re-annotated some labels to make them more closely aligned with the real-world scenarios for LLMs. For example, for Charge Prediction and Prison Term Prediction, we focus on distinguishing between similar charges and Prison Term intervals based on the legal articles. Therefore, when we

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
CrimeKgAssitant	https://github.com/LiuHC0428/LAW-GPT

Table 10: 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 syllogism, in order to obtain instruction datasets for evaluation.

re-annotated, the labeling categories and specific settings aligned with the above guidelines.

B.4 Data Description

B.4.1 Automatic Evaluation Dataset

Legal Article Recommendation. 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.

Element Recognition. 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.

Named Entity Recognition. 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.

Judicial Summarization. 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.

Case Recognition. 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.

Controversy Focus Mining. 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.

Similar Case Matching. 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.

Charge Prediction. 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.

Prison Term Prediction. 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.

Civil Trial Prediction. 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.

Legal Question Answering. 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.

Judicial Reasoning Generation. 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.

Case Understanding. 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.

Legal Consultation. 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.

B.4.2 Manual Evaluation Dataset

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.5 Data Instance

The instances of the instruction dataset are shown in Table 19-32. Specifically, the "text" field, which appears as the question in the "query" field, is omitted in the tables to avoid redundancy and save space.

B.5.1 FIR: Fundamental Information Retrieval

See Table 19-23.

B.5.2 LPI: Legal Principles Inference

See Table 24-29.

B.5.3 ALA: Advanced Legal Application

See Table 30-32.

Capability	Task	Metrics	GPT-4	ChatGPT	HanFei	wisdomInterrogatory	Fuzi-Mingcha	LexiLaw	LaWGPT	Lawyer-LLaMA	ChatLaw
FIR	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
LPI	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
ALA	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 11: 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.

C More Details of Manual Evaluation

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

C.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.

Capability	Task	Metrics	Baichuan2-Chat	Baichuan	ChatGLM	Llama-7B	Llama-13B	Llama2-Chat	Chinese-LLaMA-7B	Chinese-LLaMA-13B	Ziya-LLaMA
FIR	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
LPI	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
ALA	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 12: 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 13: 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.

- 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 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.

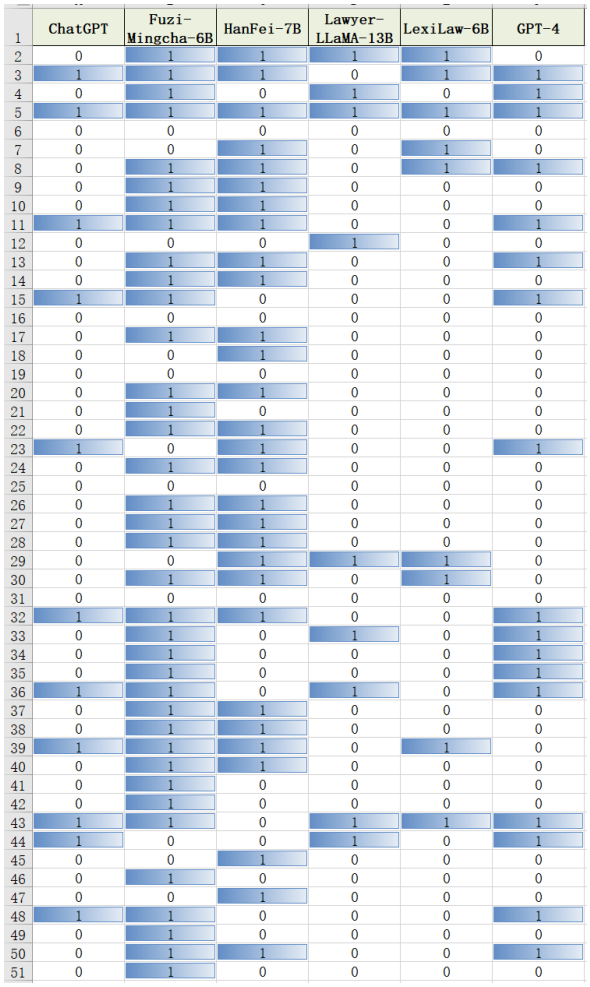
Finally, we organized the expert annotations to calculate the win rate for each LLM. Figure 2 illustrates the annotation results of expert A for the Judicial Reasoning Generation task.

C.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.

C.4 Annotators of Manual Evaluation

The three legal experts conducting the annotations are three graduate students from our research team,



	ChatGPT	Fuzi-Mingcha-6B	HanFei-7B	Lawyer-LLaMA-13B	LexiLaw-6B	GPT-4
1						
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.

specializing in the field of criminal law.

D More Details of Evaluation Metrics

For classification tasks, we select accuracy (Acc), miss rate (Miss), F1 score (F1), and matthews correlation 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 4, 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).

E 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 14: 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.

E.1 The Automatic Evaluation Results

As shown in Table 11 and Table 12, 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 syllogism, while other LLMs have strong text generation capabilities but lack syllogism.

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 syllogism.

E.2 The Win Rate of LLMs for Each Expert

As shown in Table 13, Expert A and B have similar win rates, while Expert C differs significantly from them. This suggests that while legal syllogism is commonly recognized among legal experts, there are still individual differences in actual judgment, influenced by certain subjectivity.

E.3 The Agreement Scores for Expert Evaluation

Furthermore, for the manual evaluation, we calculated agreement scores for expert evaluation, as shown in Table 14. Based on this, we observe the following fact:

Although experts can find the lack of legal syllogism in LLMs, assessing legal syllogism 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 syllogism. This allows experts to judge based on their legal syllogism, 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.

E.4 The Agreement Scores for Manual and Automatic Evaluation

We ranked the LLMs evaluated automatically based on the scores in Table 7, and ranked the LLMs evaluated manually based on the average win rate scores in Table 9. 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 16. 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.

E.5 Performance of Pre-trained LMs (PLMs)

While the evaluation of individual pre-trained language models (PLMs) for a single task doesn't capture the complete legal syllogism process, which is not our primary focus, we still explore the performance of certain PLMs in tasks similar to our benchmark, including identifying major and minor premises and drawing conclusions. This may highlight the potential of current LLMs. The results are presented in Table 15.

Capability	Task	Score	Model reference
Fundamental Information Retrieval	F1	86.13	(Yue et al., 2024)
	F2	70.95	(Zhang et al., 2022a)
	F3	-	-
	F4	-	-
	F5	-	-
Legal Princips Inference	L1	-	-
	L2	-	-
	L3	85.65	(Bi et al., 2024)
	L4	90.11	(Feng et al., 2023)
	L5	80.65	(Ma et al., 2021)
	L6	-	-
Advanced Legal Application	A1	48.17	(Wu et al., 2020)
	A2	-	-
	A3	-	-

Table 15: The results of Pre-trained LMs (PLMs) in some tasks.

Although PLMs lack the full scope of legal syllogism, they demonstrate a high sensitivity to legal characteristics, which may pave the way for LLMs. PLMs outperform LLMs significantly in tasks like classifying major and minor premises and drawing conclusions, suggesting that LLMs are less effective in these fundamental tasks compared to PLMs. It also implies that training LLMs with a focus on specific legal tasks with characteristics could enhance their relevant capabilities. However, LLMs hold the advantage of handling multiple tasks concurrently and have the potential to demonstrate the inherent structure of legal syllogism. This warrants further exploration, integrating insights gleaned from PLMs.

E.6 The In-context Results

See Tabel 17-18 for more results. The conclusions are same as section 5.2.

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 16: The agreement scores for manual and automatic evaluation.

Capability	Task	Metrics	HanFei	wisdomInterrogatory	Fuzi-Mingcha	LexiLaw	LaWGPT	Lawyer-LLaMA	ChatLaw
FIR	Legal Article Recommendation	Acc	0.2310	0.3490	0.4150	0.4150	0.1880	0.5090	0.5970
		Miss	0.0000	0.0000	0.0000	0.0000	0.6350	0.0030	0.0020
		F1	0.0867	0.2854	0.2457	0.2457	0.1996	0.4298	0.5332
	Element Recognition	Acc	0.1380	0.1990	0.0410	0.0600	0.0830	0.5590	0.5810
		Miss	0.0000	0.0000	0.7040	0.4650	0.0000	0.0000	0.0000
		F1	0.1303	0.0974	0.0280	0.0191	0.0732	0.5163	0.5783
		Mcc	0.1139	0.0785	0.0190	0.0090	0.0828	0.5312	0.5610
	Named Entity Recognition	Entity-Acc	0.2865	0.0000	0.1192	0.5317	0.0000	0.0798	0.2894
	Judicial Summarization	ROUGE-1	0.6218	0.6197	0.5774	0.4398	0.5694	0.6569	0.0206
		ROUGE-2	0.3767	0.3680	0.2939	0.2256	0.2996	0.4020	0.0092
		ROUGE-L	0.4824	0.4666	0.4235	0.3365	0.3946	0.5058	0.0157
	Case Recognition	Acc	0.8035	0.9640	0.3010	0.4890	0.5215	0.9100	0.9800
		Miss	0.0000	0.0000	0.5250	0.3250	0.0085	0.0130	0.0020
		F1	0.7958	0.9640	0.3090	0.4181	0.3837	0.9158	0.9810
LPI	Controversy Focus Mining	Acc	0.0490	0.0654	0.0621	0.0686	0.0784	0.2190	0.1340
		Miss	0.0000	0.0000	0.0752	0.3791	0.0033	0.0065	0.0261
		F1	0.0382	0.0085	0.0142	0.0314	0.0431	0.2240	0.1287
		Mcc	-0.0439	-0.0482	-0.0040	0.0371	0.0055	0.1335	0.1065
	Charge Prediction	Acc	0.4788	0.2648	0.1753	0.1717	0.0907	0.2164	0.6215
		Miss	0.0000	0.0000	0.0157	0.0000	0.4667	0.4498	0.0060
		F1	0.4401	0.1786	0.0625	0.0503	0.0803	0.1789	0.5771
	Prison Term Prediction	Acc	0.7249	0.7077	0.1289	0.0630	0.1633	0.6791	0.4699
		Miss	0.0000	0.0000	0.4642	0.5759	0.0000	0.0000	0.0000
		F1	0.6145	0.6136	0.0705	0.0405	0.0801	0.6024	0.4930
		Mcc	0.0430	0.0276	0.0471	-0.0481	0.0018	-0.0347	0.1571
	Civil Trial Prediction	Acc	0.2183	0.2120	0.2083	0.1179	0.2396	0.3149	0.2597
		Miss	0.0000	0.0063	0.6688	0.8444	0.0000	0.0000	0.0025
		F1	0.1490	0.1611	0.3114	0.1963	0.1944	0.3197	0.2293
		Mcc	0.0847	-0.0840	-0.0018	-0.0061	0.0431	0.0809	0.0800
	Legal Question Answering	Acc	0.2456	0.3392	0.2152	0.1287	0.3088	0.3240	0.3216
		Miss	0.0094	0.0000	0.2421	0.5123	0.0000	0.0012	0.0211
		F1	0.2017	0.3295	0.1581	0.1596	0.2161	0.2993	0.3134
ALA	Judicial Reasoning Generation	ROUGE-1	0.4407	0.5822	0.6393	0.2977	0.5774	0.4421	0.0642
		ROUGE-2	0.2251	0.3049	0.2655	0.1332	0.2681	0.1922	0.0279
		ROUGE-L	0.3134	0.3929	0.3740	0.2366	0.3528	0.2901	0.0570
	Case Understanding	ROUGE-1	0.8544	0.8842	0.8267	0.7639	0.8027	0.8586	0.8028
		ROUGE-2	0.8128	0.8047	0.7949	0.7041	0.7077	0.8157	0.7672
		ROUGE-L	0.8456	0.8624	0.8231	0.7539	0.7807	0.8512	0.7988
	Legal Consultation	ROUGE-1	0.3660	0.4422	0.3994	0.3201	0.2403	0.4455	0.4083
		ROUGE-2	0.1640	0.1790	0.1718	0.1269	0.0912	0.2007	0.1884
		ROUGE-L	0.2707	0.3216	0.2819	0.2350	0.1956	0.3108	0.3004

Table 17: The In-context results of Legal LLMs.

Capability	Task	Metrics	Baichuan2-Chat	Baichuan	ChatGLM	Llama-7B	Llama-13B	Llama2-Chat	Chinese-LLaMA-7B	Chinese-LLaMA-13B	Ziya-LLaMA
FIR	Legal Article Recommendation	Acc	0.7000	0.3630	0.4150	0.3670	0.2940	0.4280	0.3910	0.3400	0.6000
		Miss	0.0000	0.0000	0.0000	0.0150	0.0150	0.0000	0.1000	0.1060	0.0020
		F1	0.6595	0.2024	0.2457	0.2843	0.2197	0.3596	0.2460	0.2121	0.5808
	Element Recognition	Acc	0.7270	0.2800	0.0410	0.2870	0.3810	0.5940	0.1580	0.5070	0.5690
		Miss	0.0000	0.0000	0.7040	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		F1	0.7371	0.1723	0.0280	0.2438	0.3153	0.6136	0.1580	0.5105	0.5339
		Mcc	0.6981	0.1779	0.0190	0.1945	0.3225	0.5492	0.1659	0.4921	0.5498
	Named Entity Recognition	Entity-Acc	0.2058	0.0000	0.1192	0.0000	0.0000	0.0000	0.0000	0.3096	0.0000
	Judicial Summarization	ROUGE-1	0.5323	0.5635	0.5774	0.0014	0.0013	0.5022	0.5562	0.6455	0.0221
		ROUGE-2	0.3294	0.2854	0.2939	0.0004	0.0004	0.3019	0.3335	0.3998	0.0110
		ROUGE-L	0.4333	0.3842	0.4235	0.0009	0.0009	0.4030	0.4386	0.5057	0.0175
	Case Recognition	Acc	0.6900	0.7485	0.3010	0.5445	0.5195	0.5065	0.4995	0.6595	0.9760
		Miss	0.3090	0.0000	0.5250	0.0195	0.0350	0.0005	0.0000	0.0000	0.0060
		F1	0.7856	0.7318	0.3090	0.5296	0.4100	0.3487	0.3334	0.6195	0.9789
LPI	Controversy Focus Mining	Acc	0.2092	0.0654	0.0621	0.0654	0.0980	0.1732	0.0752	0.1111	0.1046
		Miss	0.0000	0.0000	0.0752	0.1176	0.1078	0.0033	0.0033	0.0033	0.0261
		F1	0.2265	0.0127	0.0142	0.0284	0.0297	0.1779	0.0350	0.0682	0.0840
		Mcc	0.1691	-0.0034	-0.0040	-0.0156	-0.0139	0.0652	-0.0041	0.0015	0.0430
	Charge Prediction	Acc	0.4547	0.1983	0.1753	0.1983	0.3132	0.3906	0.1753	0.1016	0.5780
		Miss	0.0024	0.0000	0.0157	0.1040	0.1100	0.0000	0.0544	0.7908	0.0060
		F1	0.4052	0.1230	0.0625	0.2031	0.3503	0.2746	0.1215	0.1601	0.5179
	Prison Term Prediction	Acc	0.5587	0.3983	0.1289	0.6791	0.6533	0.6819	0.0745	0.7249	0.6991
		Miss	0.0000	0.0000	0.4642	0.0086	0.0115	0.0000	0.0000	0.0000	0.0000
		F1	0.5823	0.4333	0.0705	0.6171	0.5878	0.6014	0.0103	0.6093	0.6154
		Mcc	0.1288	-0.0427	0.0471	0.0775	0.0127	-0.0342	0.0000	0.0000	0.0166
	Civil Trial Prediction	Acc	0.6399	0.1706	0.2083	0.1593	0.1593	0.1819	0.1669	0.1644	0.1719
		Miss	0.0000	0.0000	0.6688	0.0138	0.0125	0.0000	0.0000	0.0000	0.0025
		F1	0.6458	0.0592	0.3114	0.0456	0.0455	0.0836	0.0515	0.0465	0.0639
		Mcc	0.0871	0.0118	-0.0018	-0.0389	-0.0307	0.0370	0.0432	-0.0351	0.0286
	Legal Question Answering	Acc	0.4129	0.3076	0.2152	0.2912	0.2936	0.2889	0.2070	0.2842	0.2854
		Miss	0.0000	0.0000	0.2421	0.0000	0.0000	0.0000	0.0012	0.0000	0.0000
		F1	0.4074	0.2781	0.1581	0.1575	0.2889	0.2645	0.1094	0.2455	0.2434
ALA	Judicial Reasoning Generation	ROUGE-1	0.5831	0.5501	0.6393	0.2872	0.2407	0.3294	0.3592	0.2777	0.3683
		ROUGE-2	0.3419	0.2418	0.2655	0.1267	0.1102	0.1494	0.1506	0.1082	0.1700
		ROUGE-L	0.4561	0.3496	0.3740	0.2104	0.1961	0.2206	0.2662	0.1956	0.2583
	Case Understanding	ROUGE-1	0.8838	0.8927	0.8267	0.7215	0.7285	0.8377	0.7403	0.8740	0.8520
		ROUGE-2	0.8523	0.8167	0.7949	0.6096	0.6238	0.7606	0.6590	0.8065	0.8185
		ROUGE-L	0.8796	0.8683	0.8231	0.6906	0.6961	0.8125	0.7244	0.8530	0.8479
	Legal Consultation	ROUGE-1	0.4955	0.3606	0.3994	0.2008	0.2233	0.3988	0.2347	0.2684	0.3956
		ROUGE-2	0.2150	0.1377	0.1718	0.0671	0.0746	0.1490	0.0856	0.0933	0.1678
		ROUGE-L	0.3395	0.2721	0.2819	0.1557	0.1723	0.2669	0.1877	0.2108	0.2821

Table 18: The In-context results of baseline LLMs.

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'.

query: {**prompt**}

{**N-Shot**}

Text: The prosecution alleges: On the early morning of November 16, 2015, the defendant Zhu, together with Sun (who has already been sentenced), went to the residence of the victim Zhu in XX Village, XX Group, Pujiang Town, Minhang District, Shanghai. Zhu acted as a lookout while Sun entered through a window to commit theft, but no property was stolen. On November 18, 2015, Zhu was stopped by the police due to suspicious behavior and truthfully confessed to the above facts.

Answer:

answer: 264

choices: ["264", "133", "234"]

gold: 0

Table 19: An instance of the Legal Article Recommendation task.

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 contract 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'.

query: {**prompt**}

{**N-Shot**}

Text: After the agreement was signed, a third party brought the plaintiff and 10 others together for construction work. On August 28, 2013, Wu issued four promissory notes to the 10 plaintiffs, totaling unpaid wages of 140,070 yuan for their labor.

Answer:

answer: LB4

choices: ["LB1", "LB2", "LB3", "LB4", "LB5", "LB6", "LB7", "LB8", "LB9", "LB10", "LB11", "LB12", "LB13", "LB14", "LB15", "LB16", "LB17", "LB18", "LB19", "LB20"]

gold: 3

Table 20: An instance of the Element Recognition task.

prompt: Your task is to extract the entity 'value of the item' from the text below. If this entity does not exist, the answer is 'No'.

query: {**prompt**}

{**N-Shot**}

Text: A set of "Jingqiu" brand batteries, valued at 1488 RMB, was stolen.

Answer:

answer: 1488 RMB

Table 21: An instance of the Named Entity Recognition task.

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

query: {**prompt**}

{**N-Shot**}

Text: Plaintiff: Zhang Yinsu, male, Han ethnicity, born on March 17, 1960, residing in Jiulongpo District, Chongqing City. Authorized litigation representative: Tao Qiuyi, lawyer at Chongqing Jiuyan Law Firm. Defendant: Cai Xiaodong, male, Han ethnicity, born on December 16, 1984, residing in Wulong County, Chongqing City. This court accepted the case of the labor contract dispute between the plaintiff Zhang Yinsu and the defendant Cai Xiaodong and held a public trial according to the small claims procedure. The plaintiff Zhang Yinsu and his authorized litigation representative Tao Qiuyi attended the trial. The defendant Cai Xiaodong, having been lawfully summoned by this court, did not appear in court. The trial has now concluded. The plaintiff Zhang Yinsu has requested this court to: 1. Order the defendant to pay the plaintiff labor remuneration of 3,097 yuan for the period from August 2016 to January 2017; 2. Order the defendant to bear the litigation costs of this case. Facts and reasons: The defendant contracted the water and electricity project of the Jiaoyang Ideal City in Taojia Town, Jiulongpo District, Chongqing City. The plaintiff worked on this project from August 2016, engaging in water and electricity installation. From the beginning of the plaintiff's work until January 2017, the defendant owed the plaintiff a total wage of 3,097 yuan and issued an IOU to the plaintiff. The defendant Cai Xiaodong did not respond. After trial, this court found the following facts: The plaintiff holds a wage slip stating: "Zhang Yingshu worked in the water and electricity team at Jiaoyang Ideal City (056 plot), with a final balance of 3,097.5 yuan (in words: three thousand ninety-seven yuan and five jiao) after deductions. Please collect it from the project office by February 28, 2017." The slip is signed "Cai Xiaodong" and dated February 16, 2017. During the trial, the defendant Cai Xiaodong explained the situation to this court and acknowledged the authenticity of the wage slip and the fact that he owed the plaintiff 3,097.5 yuan in unpaid wages. This court confirms that the defendant should pay this amount. The plaintiff's claim for 3,097 yuan is respected. The plaintiff's claims are well-founded and supported. According to Article 60, Paragraph 1, and Article 109 of the Contract Law of the People's Republic of China, and Article 162 of the Civil Procedure Law of the People's Republic of China, the judgment is as follows: The defendant Cai Xiaodong shall pay the plaintiff Zhang Yinsu 3,097 yuan in labor remuneration within five days from the effective date of this judgment. If the payment is not made within the specified period, the defendant shall pay double the interest on the debt for the period of delayed performance in accordance with Article 253 of the Civil Procedure Law of the People's Republic of China. The case acceptance fee of 25 yuan shall be borne by the defendant Cai Xiaodong. This judgment is final. After this judgment takes effect, the parties must comply. If one party refuses to comply, the other party may apply for enforcement by the People's Court within two years from the last day of the specified performance period.

Answer:

answer: The plaintiff and defendant are involved in a labor contract dispute. The plaintiff requests this court to order the defendant to pay labor remuneration. The defendant did not respond. After the trial, the defendant explained to this court and acknowledged the unpaid labor remuneration owed to the plaintiff. The defendant is obligated to make the payment. According to Article 60, Paragraph 1, and Article 109 of the Contract Law of the People's Republic of China, and Article 162 of the Civil Procedure Law of the People's Republic of China, the court orders the defendant to pay the plaintiff labor remuneration.

Table 22: An instance of the Judicial Summarization task.

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.

query: {**prompt**}

{**N-Shot**}

Text: Upon trial, it was found that the defendant, Li, is the actual operator of Changsha Kaicheng Paper Co., Ltd. In the second half of 2013, Li was introduced by Yi to the defendant Tan, the deputy branch manager of the Dongtang Branch of China Construction Bank, with the intention of obtaining a credit loan of 65 million yuan from China Construction Bank for Changsha Kaicheng Paper Co., Ltd. In early November 2013, Tan officially started the credit approval process, and through Tan's operations, the 65 million yuan credit approval for Changsha Kaicheng Paper Co., Ltd. was successfully granted by the Hunan Province Branch of China Construction Bank on December 31, 2013. On January 4, 2014, to thank Tan for his help, Li withdrew 300,000 yuan from the account of Zhang, the legal representative of Changsha Kaicheng Paper Co., Ltd., and delivered the 300,000 yuan in cash to Tan at Lu's Tea House near Xiangfu Huacheng, Tianxin District. After receiving the money, Tan immediately deposited the 300,000 yuan into an account ending in 0977, held by his mother Cheng at China Construction Bank, and later used it to purchase a financial product.

Answer:

answer: Criminal

choices: ['Civil', 'Criminal']

gold: 1

Table 23: An instance of the Case Recognition task.

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'.

query: {**prompt**}

{**N-Shot**}

Text: PER appeals, stating that after marriage, Wang and PER frequently quarreled due to personality and other differences. Currently, there is no affection between them, and their marriage is in name only, making it impossible to live together. The original judgment incorrectly concluded that the marital relationship had not completely broken down. PER requests the appellate court to revise the judgment according to the law. PER argues that the marital relationship has not broken down and does not agree to the divorce.

Answer:

answer: F

choices: ["A", "B", "C", "D", "E", "F", "G", "H", "I", "J"]

gold: 5

Table 24: An instance of the Controversy Focus Mining task.

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.

query: {**prompt**}

Text: 'A': Plaintiff Qin Yuanhuo, unemployed. Defendant Kong Zhimin, businessman. The plaintiff, Qin Yuanhuo, claims that on September 25, 2014, he mortgaged his residential property certificate at the Dongxing City Credit Union Bank for a loan of 500,000 yuan, which he lent to the defendant, Kong Zhimin. It was agreed that if the defendant did not repay the loan on time, he would be in breach of contract, and all costs incurred by the plaintiff to realize the debt, including court fees and enforcement fees, would be borne by the defendant. Due to 200,000 yuan not yet being due, the plaintiff withdrew that portion and changed the claim to 300,000 yuan. The lawsuit requests are: 1. The defendant repays the plaintiff 300,000 yuan and interest of 28,000 yuan (calculated at 3,500 yuan per month for 8 months); 2. The defendant bears the litigation costs. The main evidence provided by the plaintiff within the evidence submission period includes: 1. ID card, proving the plaintiff's identity; 2. IOU, proving that the defendant borrowed 500,000 yuan from the plaintiff. The defendant, Kong Zhimin, responds that he has no objection to the plaintiff's claim and will repay the loan according to the IOU, aiming to clear the debt by the end of 2016. The defendant did not provide any evidence within the evidence submission period. After a hearing, the defendant had no objection to the plaintiff's evidence items 1 and 2. This court confirms the evidence to which neither party has objected. After the trial, it was found that the defendant borrowed money from the plaintiff for business purposes and issued an IOU on September 25, 2014, stating: "Borrowed from Qin Yuanhuo, 500,000 yuan. Interest calculated according to the bank's rate, being 3,500 yuan per month. Repayment schedule: 100,000 yuan on December 30, 2014; 200,000 yuan on December 30, 2015; full repayment of 500,000 yuan on December 30, 2016. Interest settled monthly by the 30th." The defendant failed to repay the loan on time.

'B': Plaintiff Liang, residing in Chongxin County, Gansu Province. Defendant Du, residing in Chongxin County. The plaintiff, Liang, filed a lawsuit requesting: 1. The defendant immediately repay the plaintiff 310,000 yuan and interest of 200,000 yuan; 2. The defendant bears the litigation costs. Facts and reasons: The plaintiff and the defendant are friends. On March 26, 2012, the defendant borrowed 100,000 yuan from the plaintiff for a coal business, followed by additional loans of 70,000 yuan on April 19, 2012, 70,000 yuan on May 2, 2012, and 70,000 yuan on July 20, 2012. The agreed interest rate was 4% per month. Initially, the defendant repaid some interest, approximately 140,000-150,000 yuan, but no payments were made after early 2014. The plaintiff could not locate the defendant after October 2015. The defendant did not appear in court nor respond. The court found that the plaintiff and the defendant were middle school classmates and friends. The defendant borrowed money due to insufficient funds for a coal supply partnership with Chongxin Power Plant, borrowing a total of 310,000 yuan on the mentioned dates with a verbal agreement of 4% monthly interest. The defendant repaid approximately 150,000 yuan in interest before early 2014 but failed to pay interest or principal thereafter. The plaintiff's repeated demands were met with promises of repayment upon obtaining a loan, which the defendant never secured. The defendant resigned from his position in May 2016. The plaintiff filed the lawsuit on February 28, 2017. Evidence includes four original IOUs and the response from Chongxin County Water Bureau and Human Resources and Social Security Bureau regarding the defendant's resignation.

'C': Plaintiff Zheng Zhihua, male, Han ethnicity, residing in the Mining District of Datong City. Defendant Wu Tongseng, male, Han ethnicity, residing in the Urban District of Datong City. The plaintiff, Zheng Zhihua, filed a lawsuit requesting: 1. The defendant repay 10,000 yuan and interest of 2,880 yuan up to July 2016, totaling 12,880 yuan; 2. The defendant bears the litigation costs. Facts and reasons: On June 15, 2015, the defendant pleaded with the plaintiff for 10,000 yuan needed to complete his retirement process, promising to repay the principal and interest upon receiving his pension. The plaintiff initially refused but eventually agreed after repeated requests. The plaintiff borrowed 10,000 yuan from a colleague and lent it to the defendant, who promised to repay within two months and issued an IOU. Despite receiving his pension, the defendant failed to repay after more than a year. To protect his contractual rights, the plaintiff filed this lawsuit. The defendant acknowledged the loan and agreed to repay the 10,000 yuan principal but disputed the interest, stating that the initial agreement included a 3% interest rate, and the plaintiff only gave him 9,000 yuan after deducting 1,000 yuan interest upfront. The court found that the defendant borrowed 10,000 yuan from the plaintiff on June 15, 2015, and issued an IOU stating: "Borrowed 10,000 yuan from Brother Zheng, at 3% interest, to be repaid in about two months." The defendant signed the IOU. The court did not accept the defendant's claim that he should not pay interest, as the IOU specified both the interest rate and repayment period. The defendant's claim of an upfront deduction of 1,000 yuan interest was also not supported by evidence.

Answer:

answer: B

choices: ["B", "C"]

gold: 0

Table 25: An instance of the Similar Case Matching task.

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'.

query: {**prompt**}

{**N-Shot**}

Text: The prosecution alleges that from April 5, 2016, to April 14, 2016, the defendants Cao and Li conspired to commit six thefts in various residential areas in Yangxin County, Binzhou City, Wulian County, Rizhao City, and Yishui County, using technical unlocking methods to steal cash, gold jewelry, and other items valued at over 7,600 yuan. The specific facts of the crimes are as follows: On April 5, 2016, at around 1 PM, the defendants Cao and Li went to a residential area in Yangxin County, Binzhou City. Li knocked on the door and kept watch while Cao unlocked the door and entered the house, stealing a gold ring, a pair of gold earrings, a Lenovo tablet, a charger, and a 500GB external hard drive from Meng's home, totaling over 4,700 yuan. After the incident, the pair of gold earrings was returned to the victim, Meng. On April 12, 2016, at around 4 PM, the defendants went to a residential area in Wulian County, Rizhao City. Li knocked on the door and kept watch while Cao unlocked the door and entered the house, stealing a gold ring and over 600 yuan in cash from Xu's home, totaling over 2,800 yuan. On April 14, 2016, the defendants went to a residential area in Xujiahu Town, Yishui County. Li knocked on the door and kept watch while Cao unlocked the door and entered the houses of residents Zhang, Wang, Gao, and Du. They stole ten 10-yuan bills with consecutive serial numbers from Wang's home in Building 1 and over 70 yuan in cash from Gao's home in Building 2, totaling over 170 yuan. They were caught by Du Yuwei while stealing from Du's home, who then called the police. The above facts were not disputed by the defendants during the trial and are corroborated by on-site inspections, examination records, identification records, physical evidence and photographs, documentary evidence, statements from victims such as Meng, and the confessions of defendants Cao and Li, which are sufficient to establish the facts.

Answer:

answer: 69

choices: ["69", "50", "124"]

gold: 0

Table 26: An instance of the Charge Prediction task.

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'.

query: {**prompt**}

{**N-Shot**}

Text: The Zhuji City People's Procuratorate alleges that on September 21, 2012, at around 2 PM, in Zhuzhong Village, Huandong Subdistrict, Zhuji City, Shou and Guo Fang had a dispute and a physical altercation over garbage disposal at the entrance of their homes. During this time, Yuan Guohong and Feng went to intervene. The defendant, Yuan, believing that the victim Feng was forcibly intervening, went to Guo Fang's home with a hoe and struck Feng on the right shoulder and other areas, causing minor injuries. To prove the above accusations, the prosecution has provided corresponding evidence to the court, asserting that the defendant, Yuan, intentionally injured another person, resulting in minor injuries, and should be held criminally responsible for intentional injury. The prosecution requests that the court punish the defendant according to Article 234, Paragraph 1 of the Criminal Law of the People's Republic of China.

Answer:

answer: B

choices: ["A", "B", "C"]

gold: 1

Table 27: An instance of the Prison Term Prediction task.

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.

query: {**prompt**}

{**N-Shot**}

The facts are as follows: The two defendants are spouses. Defendant PER borrowed money from the plaintiff intermittently from March 1, 2011, to July 30, 2011. On January 13, 2012, both parties settled the previous loans, with PER owing the plaintiff a remaining sum of 400,000 yuan. PER personally issued a promissory note, agreeing to repay the amount by February 13, 2012. As of now, the aforementioned loan remains unpaid, leading to this litigation.

The plaintiff's claim is as follows: Request the court to order the defendants to jointly repay the plaintiff's loan of 400,000 yuan and the interest loss (calculated based on the ORG's comparable loan prime rate from February 14, 2012, until the date of repayment determined by the judgment).

Answer:

answer: A

choices: ["A", "B", "C"]

gold: 0

Table 28: An instance of the Civil Trial Prediction task.

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.

query: {**prompt**}

{**N-Shot**}

Question: Which of the following matters creates legal relations regulated by labor law?

'A': Chen, an employee of Company A, accidentally injures customer Jiang while operating a machine during a visit. Jiang demands compensation from Chen.

'B': Wang is rejected for employment by Company B due to being under 1.70 meters in height, and intends to sue Company B.

'C': Company C reaches an agreement with a labor service company regarding 30 dispatched laborers.

'D': Company D provides a guarantee to a bank for its employees' housing purchases.

Answer:

answer: B

choices: ["A", "B", "C", "D"]

gold: 1

Table 29: An instance of the Legal Question Answering task.

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

query: {**prompt**}

{**N-Shot**}

['Claims': 1. The defendant shall repay the plaintiff's loan of 200,000 yuan and compensate for interest losses (calculated based on the benchmark interest rate for loans of the same period and grade published by the People's Bank of China from September 14, 2012, until the repayment is completed). 2. The defendant shall bear the litigation costs of this case.

'Findings': After the trial, the court established the following facts: On September 7, 2012, the defendant, due to the need for capital turnover, borrowed 200,000 yuan from the plaintiff and issued an IOU on the same day, agreeing to repay the loan by September 13, 2012. After the loan became due and the plaintiff's demands for repayment, the defendant has yet to repay the loan.]

The court holds that:

answer: The private loan relationship between the plaintiff and the defendant is legal and valid and should be protected by law. The defendant's failure to repay the 200,000 yuan loan to the plaintiff is clearly established and supported by sufficient evidence. The defendant is therefore civilly liable to repay the loan and compensate for interest losses. Consequently, the plaintiff's claims are legitimate and supported by evidence, and this court upholds them. The defendant, having been legally summoned by this court, refused to appear in court without a valid reason, showing a lack of respect for the law and a waiver of their litigation rights. The resulting legal consequences shall be borne by the defendant.

Table 30: An instance of the Judicial Reasoning Generation task.

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".

query: {**prompt**}

{**N-Shot**}

'Legal text material': The prosecution alleges that between May 31 and June 8, 2013, the defendant Fei, in collusion with Hu1 and Zhou2, illegally cut down 321 trees, totaling 41.7244 cubic meters of timber, in Linban 151, Section 2, Daobao Village, Daobao Town, Taobei District, Baicheng City, without obtaining a logging permit. Between July 15 and 22, 2013, the defendant Fei, together with Zhou2, illegally cut down poplar trees belonging to villagers Qu and others in Daobao Town, totaling 54.1825 cubic meters of timber. Between March and April 2013, the defendant Fei, in collusion with Li, defrauded Luo of 50,000 yuan under the pretext of matchmaking. The prosecution argues that the defendants Fei, Zhou2, and Hu1 violated national forest protection regulations by illegally logging without approval from the forestry administrative department, with the quantity of logged trees being significant; and by illegally cutting down privately owned trees in large quantities, thereby violating Article 345, Paragraphs 1 and 2 of the Criminal Law of the People's Republic of China. The facts of the crime are clear, and the evidence is sufficient to hold the defendants Fei and Zhou2 criminally responsible for the crimes of illegal logging and theft of trees, and the defendant Hu1 criminally responsible for the crime of illegal logging. The defendants Fei and Li, with the intent of illegal possession, defrauded others of 50,000 yuan under the pretext of a fraudulent marriage and dowry, which constitutes a large amount, thereby violating Article 266 of the Criminal Law of the People's Republic of China. The facts of the crime are clear, and the evidence is sufficient to hold the defendants Fei and Li criminally responsible for the crime of fraud.

'Question': What is the volume of poplar trees illegally cut by the defendant Fei and others in Bao Town Village?

Answer:

answer: 54.1825 cubic meters

Table 31: An instance of the Case Understanding task.

prompt: If you are a lawyer, please answer the legal consultation question below based on the real scenario.
query: {**prompt**} {**N-Shot**} Question: Will there be any property loss in the event of a divorce for properties owned before obtaining the marriage certificate? Answer:
answer: If there is a prenuptial agreement or a property division agreement specifying that the property owned before obtaining the marriage certificate belongs to one party, there will be no property loss in the event of a divorce. However, without such an agreement, the property will be considered marital property and will need to be divided according to relevant laws. It is important to note that property that existed before the marriage and appreciated in value during the marriage will also be considered marital property and subject to division. Therefore, if you own property, it is advisable to create a prenuptial agreement or a property division agreement before marriage to avoid potential property loss.

Table 32: An instance of the Legal Consultation task.