

UCL-Bench: A Chinese User-Centric Legal Benchmark for Large Language Models

Ruoli Gan^{a,*}, Duanyu Feng^{b,*}, Chen Zhang^c, Zhihang Lin^a, Haochen Jia^b,
Hao Wang^b, Zhenyang Cai^a, Lei Cui^e, Qianqian Xie^f, Jimin Huang^{f,†},
Benyou Wang^{a,†}

^aThe Chinese University of Hong Kong, Shenzhen, ^bSichuan University,
^cNational University of Singapore, ^dGuizhou Communications Polytechnic University,
^eThe Fin AI
jimmin.huang@thefin.ai wangbenyou@cuhk.edu.cn

Abstract

Existing legal benchmarks focusing on knowledge and logic effectively evaluate LLMs on various tasks in legal domain. However, few have explored the practical application of LLMs by actual users. To further assess whether LLMs meet the specific needs of legal practitioners in real-world scenarios, we introduce UCL-Bench, a Chinese User-Centric Legal Benchmark, comprising 22 tasks across 5 distinct legal scenarios. To build the UCL-Bench, we conduct a user survey targeting legal professionals to understand their needs and challenges. Based on the survey results, we craft tasks, verified by legal professionals, and categorized them according to Bloom's taxonomy. Each task in UCL-Bench mirrors real-world legal scenarios, and instead of relying on pre-defined answers, legal experts provide detailed answer guidance for each task, incorporating both "information" and "needs" elements to mimic the complexities of legal practice. With the guidance, we use GPT-4 as the user simulator and evaluator, enabling multi-turn dialogues as a answer guidance based evaluation framework. Our findings reveal that many recent open-source general models achieve the highest performance, suggesting that they are well-suited to address the needs of legal practitioners. However, these legal LLMs do not outperform ChatGPT, indicating a need for training strategies aligned with users' needs. Furthermore, we find that the most effective models are able to address legal issues within fewer dialogue turns, highlighting the importance of concise and accurate responses in achieving high performance. The code and dataset are available at <https://github.com/wittenberg11/UCL-bench>.

1 Introduction

The legal field has shown a strong interest in large language models (LLMs), seeing their potential to

offer support to professionals in law firms, courts, and prosecutorial offices, while also providing legal consultations to the general public (Kuppa et al., 2023; Cheong et al., 2024; Murray, 2023). To facilitate a better understanding and evaluation of LLMs within the legal field, it is crucial to develop benchmarks that assess various aspects, such as legal knowledge and legal reasoning abilities (Huang and Chang, 2022; Burton, 2017).

Several legal benchmarks have been developed to address this need. For example, LegalBench, an English benchmark developed in partnership with legal experts, features 162 tasks spanning six types of reasoning (Guha et al., 2023). Similarly, LAiW and Lawbench have been proposed for the Chinese legal system (Dai et al., 2023; Fei et al., 2023). However, existing legal benchmarks still have two shortcomings: (1) *Current benchmarks fail to directly assess LLM performance from the user perspective, which may create a gap between evaluation results and real-world performance.* Existing legal benchmarks primarily test legal knowledge and logical reasoning, but they fail to directly address the specific needs and practical operations of various legal users in real scenarios (Guha et al., 2023; Fei et al., 2023). For example, lawyers may need comprehensive case analysis and strategy development, judges require balanced and fair legal interpretations, prosecutors focus on gathering and presenting evidence, while law professors emphasize thorough research and teaching (Schultz, 1992; Leeuw and Schmeets, 2016). Current benchmarks do not fully capture these varied demands, leading to a gap between benchmark results and actual user experiences. (2) *Current benchmarks fail to capture dynamic characteristic of User-LLM interactions.* Most benchmarks use a static testing approach, with many tests consisting of multiple-choice or short-answer questions (Dai et al., 2023). This method fails to mimic real user-LLM interactions, where users' queries are often conversa-

*Co-first Authors.

†Corresponding Authors.

tional and involve multi-turn exchanges (Liu et al., 2023a,b). As a result, the existing benchmarks do not directly evaluate the model’s capabilities from the perspective of user experience, which may result in a gap between the model’s capabilities from the evaluation and its actual performance in real-world scenarios.

Therefore, we introduce UCL-bench, a user-centric chinese law benchmark, with approach shown in Figure 1. We first conduct user surveys before constructing the dataset to ensure that our benchmark reflects real user needs. Through comprehensive user surveys across diverse legal scenarios, we and legal experts identify 22 tasks under the Bloom taxonomy, covering high-frequency real-world legal use cases of LLMs. Each data instance in UCL-bench consists of ‘information-needs’ pair, reflects distinct user requests and is accompanied by expert-annotated answer guidance, specifying the content required for high-quality responses. For the evaluation framework, to simulate real-world User-LLM interaction, a user-simulator (GPT-4) is employed to facilitate multi-round interactions with LLMs, based on "information" and "needs" in each sample. We further employ LLM-as-a-Judge paradigm (Zheng et al., 2024) to evaluate the service provided by LLMs with the sample-granularity answer guidance.

The results of our benchmark reveals three main findings. Firstly, as the highest performance OpenAI o1-preview and Qwen-2.5-7B-Instruct achieved, we highlight the effectiveness of recent open-source general models in user-centric legal contexts. Secondly, despite pre-trained and fine-tuned with legal corpora, legal LLMs did not surpass ChatGPT, necessitates training strategies that are better aligned with user needs. Thirdly, we find that more effective models tend to address legal issues with fewer dialogue turns, suggesting that concise and accurate responses are a key factor in achieving high performance. We also conducted additional experiments to validate the reliability of our evaluation framework.

Our contributions are threefold:

- We introduce User-Driven Legal Benchmarking Dataset covering high-frequency real-world legal use cases of LLMs. Through comprehensive user surveys across diverse Chinese legal scenarios, we identified these tasks, featuring "information-needs" pairs for distinct user requests and expert-annotated an-

swer guidance for high-quality responses.

- We introduce a novel user-centric evaluation framework for LLMs in legal scenarios. This framework leverages a GPT-4-powered user simulator to mimic real-world interactions and employs expert-guided automatic evaluation with GPT-4 as a judge, providing a more practical assessment for LLMs.
- Our benchmark reveals that open-source general models excel in user-centric legal contexts, highlighting the need for training strategies that prioritize user needs and concise, accurate responses for high performance.

2 Related Work

2.1 Legal Benchmark

With LLMs showing promising legal abilities (Huang et al., 2023; Cui et al., 2023), thoroughly evaluating their skills in this domain becomes vital. This not only aids in exploring the boundaries of capabilities of LLMs but also allows for the design of applications to better assist various users in the legal field. Currently, the construction of legal benchmarks is primarily based on two approaches: knowledge-based and legal reasoning logic-based. In the knowledge-based category, except solely constructing benchmark from the NLP tasks (Niklaus et al., 2023), many legal benchmarks construct for legal understanding (Hwang et al., 2022; Chalkidis et al., 2022; Joshi et al., 2024), and LawBench (Fei et al., 2023) first constructs a legal benchmark based on memorization, understanding, and application of legal knowledge. DISC-Law-Eval (Yue et al., 2023) divides evaluation into objective and subjective parts, using automated metrics or ChatGPT to assess legal knowledge memorization and application abilities. The legal reasoning logic-based category includes LAiW (Dai et al., 2023) and LEGALBENCH (Guha et al., 2024), which are based on the syllogism/IRAC logic of jurisprudence, and assess different logical levels in Chinese and English respectively. Although these benchmarks successfully evaluate LLMs’ knowledge and logic, they overlook the varied uses of LLMs by different users. A evaluation framework catering to diverse users is essential for different needs of users. Moreover, due to data collection and evaluation challenges, these benchmarks miss out on multi-turn dialogue evaluations, crucial for assessing LLMs in realistic legal scenarios.

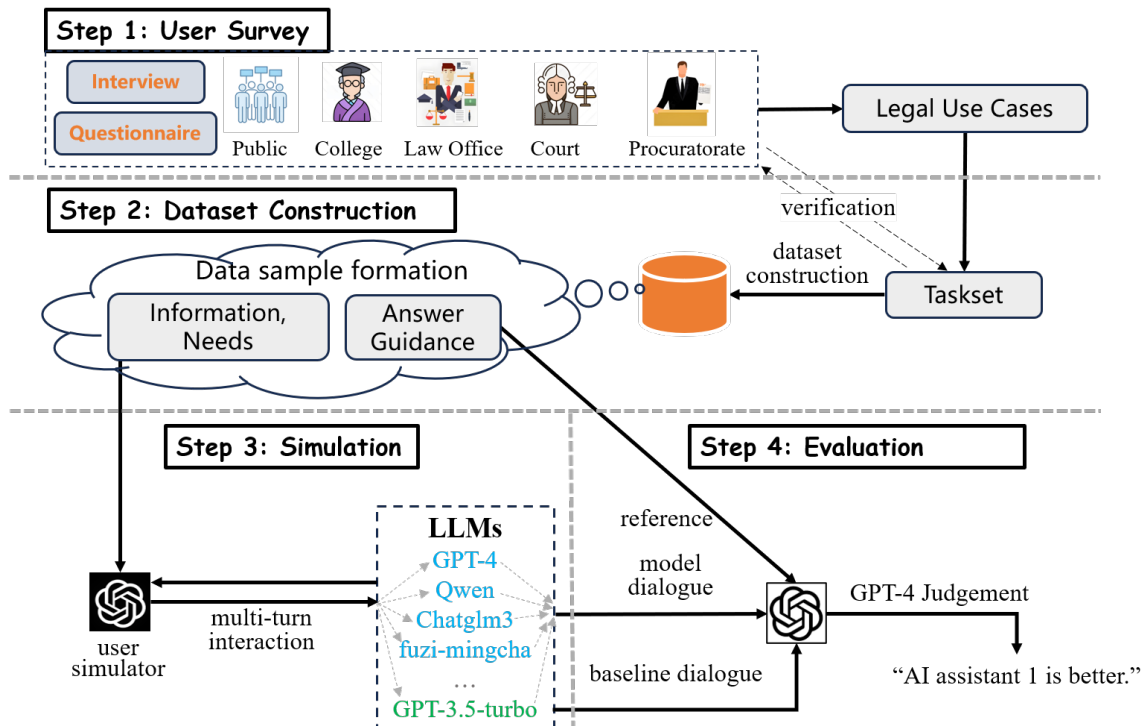


Figure 1: UCL-Bench takes a user-centric approach to evaluating LLMs in legal domain. It identifies real-world legal use cases through user surveys and then constructs tasks based on these needs. Each task is built from open-source data and public websites, incorporating both "information" and "needs" elements to mimic user requests. To ensure rigorous evaluation, expert-annotated "answer guidance" is provided for each task, including both ground truth and explanations. GPT-4 (Achiam et al., 2023) simulates real-world user interactions with different LLMs, capturing multi-turn dialogues. These dialogues are then assessed by another GPT-4, using ChatGPT (Roumeliotis and Tselikas, 2023) as a baseline and expert-annotated guidance as a reference point.

2.2 Other Benchmarks for LLM

Recently, many benchmarks have been developed to test the limits of LLMs (Chang et al., 2023), primarily using fixed instructions to assess their capabilities. For instance, AlpacaEval (Li et al., 2023) focuses on evaluating the ability of LLMs to generate appropriate responses given a diverse set of questions, MMLU (Hendrycks et al., 2020) focuses on the effective application of knowledge across various subjects, and Big-Bench (Srivastava et al., 2022) introduces even more comprehensive tasks that are believed to be beyond the capabilities of current language models. While these benchmarks offer detailed assessments of LLMs across various dimensions, they fall short in probing LLM capabilities in complex, multi-turn dialogue scenarios. Currently, the most popular general benchmark for multi-turn dialogues is MT-bench (Zheng et al., 2024), along with its derivatives (Sun et al., 2023b; Bai et al., 2024). These benchmarks use predefined questions, responses, and a fixed number of dialogue turns for each sample. During testing, each dialogue turn is input into the LLM to assess its

response generation capabilities. The fixed number of turns and predetermined questions constrain LLMs' response flexibility, diverging from real-life scenarios. Yet, high-quality multi-turn dialogue benchmarks in legal domain even general domain are still lacking.

3 UCL-bench

This section details the methodology employed for the development of our user-centric law benchmark. We commence with a user survey conducted according to established protocols, aimed at identifying critical tasks for benchmark construction. Utilizing the results of the survey, we subsequently construct a comprehensive dataset for our benchmark. The section also concludes with a description of our novel evaluation methodology, which more closely reflects real-world interactions.

3.1 User Survey

We conduct comprehensive user surveys to identify the real world Chinese legal scenarios and use cases where LLMs can offer assistance.

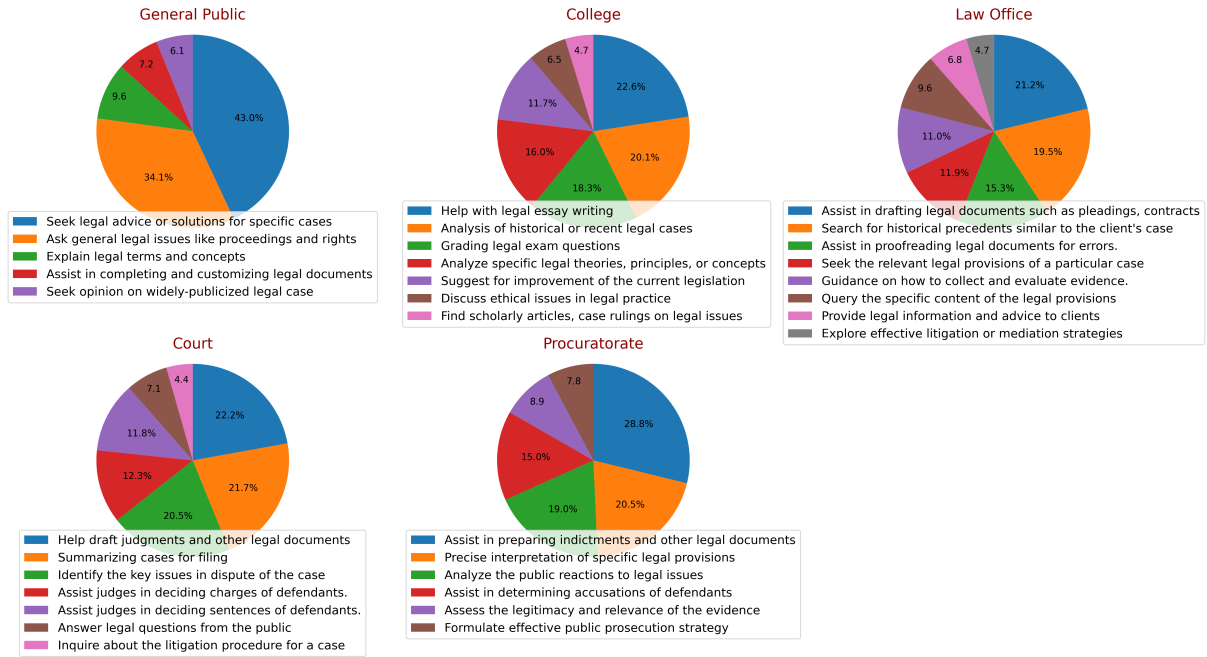


Figure 2: Proportions of use cases mentioned in every scenario. A total of 33 use cases were collected, with 28 from semi-structured interviews and 5 from open questions in questionnaires. The questionnaire results were divided according to legal scenarios, and the proportion of use cases was calculated respectively.

Semi-structured interviews (Kakilla, 2021) were conducted with eleven legal experts, including three law professors, two legal LLM expert, two senior attorney, two judges and two procurators. Demographic background of the participants and the interview protocol are shown in Appendix A.1. These conversations enabled us to collect a range of legal scenarios and potential use cases of LLM in these scenario. As a result, we determined 5 scenarios including **General public**, **College**, **Law office**, **Court**, and **Procuratorate**, and 28 use cases spanning the five scenarios. These use cases serve as candidates for subsequent questionnaire voting, see the details in Appendix A.2.

Then, we developed an online questionnaire to understand which use cases legal users are more interested in. Participants were first asked to describe their specific scenarios. This was followed by detailed inquiries into the use cases that interested them within those scenarios, using the previously collected use cases as candidates, along with an open-ended question for proposing any additional use cases. Details can be found Appendix A.2. After the responses were collected, we calculated and ranked the proportions of each use case being selected or mentioned. The results are shown in Figure 2. Further, we excluded use cases with a proportion lower than 10% as low-frequency applications, ultimately retaining 22 use cases, see

Table 5. These retained use cases reflect the high-frequency potential legal applications of LLMs.

3.2 Dataset Construction

1. Task Formulation through Experts Verification. We had legal experts design tasks for each use case. These tasks were validated through multiple rounds of feedback questionnaires, allowing us to make necessary revisions, see Appendix B.1 for details. We categorize these tasks into **Memory**, **Understand**, **Apply**, **Analyze**, **Evaluate**, and **Create**, according to Bloom’s cognitive levels (Krathwohl, 2002), see Appendix B.2 for more introduction. The categorized tasks and descriptions are shown in Appendix B.3.

2. Data Collection and Annotation. Statistics of our dataset are shown in Table 1. Figure 3 shows an instance of our dataset. Each sample is organized into "information-needs" pair. "Information" reflects the legal context before user querying the LLM, and "needs" specify the user’s requirements. These contents are used as prompt for the user-simulator to dynamically generate questions. The data sources of the "information-needs" pair are from Chinese public legal websites or open-source datasets. Refer to Table 7 in the Appendix B.3 for the corresponding licenses and data proportion. We collect the materials and reformatted them, with specific reformatting approaches by legal experts

Cognitive Level	ID	Task Name	Data Source	n
Memory	1-1	Article Recitation	flk.npc.gov.cn	50
	2-1	Legal Issue Consult	imlaw.cn	50
Understand	2-2	Judicial Interpret	flk.npc.gov.cn	50
	2-3	Legal Concept Und.	CAIL2022	50
	2-4	Public Opinion Summary	CAIL2022	50
	2-5	Judicial Summary	CAIL2021	50
	3-1	Prison Term Prediction	CAIL2021	50
Apply	3-2	Charge Prediction	wenshu.court.gov.cn	50
	3-3	Accusation Prediction	12309.gov.cn	50
	4-1	Article Prediction	wenshu.court.gov.cn	50
Analyze	4-2	Dispute Focus Identify	wenshu.court.gov.cn	50
	4-3	Real Case Consult	12348.gov.cn	50
	4-4	Similar Case Retrieval	LeCaRD (Ma et al., 2021)	50
	4-5	Legal Case Analysis	Chinese judicial exam	50
	5-1	Judicial Exam Score	Chinese judicial exam	50
Evaluate	5-2	Judgment Doc. Proofread	wenshu.court.gov.cn	50
	6-1	Legal Essay Writing	Chinese judicial exam	50
Create	6-2	Cross-Examination Gen.	wenshu.court.gov.cn	50
	6-3	Litigation Strategy Gen.	moj.gov.cn	50
	6-4	Judgment Doc Gen.	wenshu.court.gov.cn	50
	6-5	Procuratorial Doc Gen.	12309.gov.cn	50
	6-6	Lawyer Document Gen.	self-construct	50

Table 1: Statistics of Task Set. 22 tasks are classified based on cognitive levels. Most of the data come from the publications on the official website, others are come from open-sourced dataset like Challenge of AI in Law(CAIL) dataset. The total amount of data is 1100.

for each task also detailed in Appendix B.3.

In addition, experts provide fine-grained annotations for each sample, termed as "Answer Guidance". The Answer Guidance include not only the ground truth but also explanatory content, defining the standards of professionalism for the answers that LLM should provide. The four components of answer guidance are: (1) Ground Truth, (2) Mandatory Mention, (3) Advisable Mention, (4) Encouraged Mention. See Table 2 for explanation.

Category	Content
Ground Truth	Precise answers or solutions.
Mandatory Mention	Critical points or facts.
Advisable Mention	An explanation of a relevant topic or concept.
Encouraged Mention	Relevant reference, e.g., legal provision.

Table 2: Components of Expert-annotated Answer Guidance.

3.3 User-LLM Interaction Simulation

UCL-Bench distinguishes itself from existing legal benchmarks by incorporating user-LLM interaction simulation, as depicted in Figure 4. This is crucial for our tasks, which involve utilizing LLMs to solve real-world legal problems through multi-turn dialogues and simulating multi-turn user interactions.

Specifically, GPT-4 is employed as the user simulator to engage in multi-turn dialogues with the LLM under evaluation, which demonstrates strong capabilities in diverse role-playing scenarios (Wang et al., 2023). The prompt template of the user-simulator is shown in Figure 9. Through this ap-

proach, we simulate real-world LLM usage scenario in legal field. Appendix F shows dialogue examples of each task.

3.4 Evaluation

To evaluate the performance of different LLMs, we utilize the LLM-as-a-Judge paradigm (Zheng et al., 2024) by employing GPT-4 as the evaluator. Similar to AlpacaEval (Li et al., 2023), we conduct pairwise comparisons of multi-turn interactions produced by the candidate LLM and a reference LLM. ChatGPT serves as the reference LLM and the performance of other models is quantified by their win rate over the reference LLM. However, unlike existing evaluation frameworks, considering that GPT-4 evaluator may lack the necessary expertise or exhibit biases, we include answer guidance as references to the GPT-4 evaluator, ensuring GPT-4 has the necessary background knowledge when making judgments.¹

Furthermore, our evaluation approach also differs from existing methods used for evaluating dialogue systems with simulators (Sun et al., 2023a). Unlike dialogue systems with simulators, which rely on fixed overall metrics like task success rate and further evaluate each sentence (Deriu et al., 2021), we evaluate the entire conversation after it concludes using GPT-4 with answer guidance. This

¹The instruction template for prompting the GPT-4 evaluator is in Appendix D.

<p>Information: Ma Wu built a house without obtaining a construction project planning permit. On March 12, 2022, the Urban Planning Bureau issued a "Notice of Order to Demolish Illegal Construction" to Ma Wu, ordering him to demolish it within one day. On March 15, 2022, the Port City Task Force organized a forced demolition operation and demolished Ma Wu's illegal construction. During the demolition, Ma Wu had not yet had time to move out the items inside the house. The Port City Task Force also did not legally register and preserve the items inside, nor did they make an inventory list and have Ma Wu sign to confirm it. It was later discovered that the Urban Planning Bureau had sent a letter of authorization to the Port City Task Force, authorizing them to make the decision to demolish the illegal construction.</p> <p>Needs: Want to know if the conduct of the Urban Planning Bureau is legal.</p> <p>Answer Guidance:</p> <ol style="list-style-type: none"> 1. Ground Truth: Illegal. 2. Mandatory Mention: Violation by the responsible entity; Procedural violation; Violation of time limits 3. Advisable Mention: None 4. Encouraged Mention: Article 68 of the Urban and Rural Planning Law; Articles 35, 36, and 44 of the Administrative Enforcement Law
--

Figure 3: A data instance in UCL-bench. The corresponding task is Legal Case Analysis, requiring LLM to analyze legal issues according to the context of the case.

approach prioritizes flexibility, as we are more concerned with whether the LLM successfully resolves the legal issue through the conversation, rather than requiring perfection in every detail of the dialogue.²

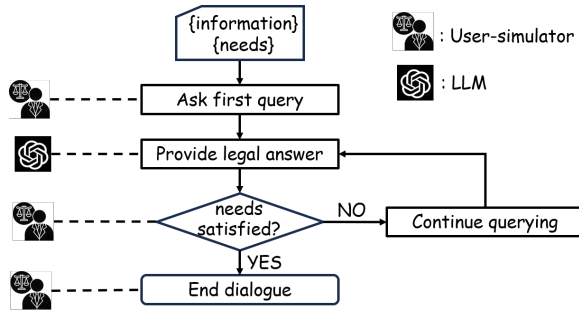


Figure 4: The process of multi-turn dialogue generation between user simulator and LLM under evaluation. "Information-needs" pair serve as prompting input to the user simulator. The simulator then formulates a query and determines whether to continue or terminate the dialogue based on the responses it gathers. The maximum number of simulator-LLM exchanges is limited to three (6 turns in total).

4 Experiment

4.1 Setup

We selected 21 models for evaluation, including 4 closed-source LLMs: ChatGPT, GPT-4, OpenAI o1-mini and OpenAI o1-preview. 14 open-source general LLMs: Baichuan2-7B-Chat, Baichuan2-13B-Chat (Baichuan, 2023), chatglm3-6b (Du et al., 2022), deepseek-llm-67b-chat (DeepSeek-AI et al., 2024), Qwen-7B-Chat, Qwen-14B-Chat, Qwen-72B-Chat, Qwen-2.5-7B-Instruct (Bai et al., 2023), internlm-chat-7b, internlm-chat-20b, internlm-2.5-7B-Chat (Cai et al., 2024), Yi-34B-Chat (AI

et al., 2024), DeepSeek-V2-Chat (Bi et al., 2024), GLM-4-9B-Chat (GLM et al., 2024), and 3 legal-specific models: Wisdom-Interrogatory (Yiquan et al.), DISC-LawLLM (Yue et al., 2023), fuzimingcha (Wu et al., 2023). See Appendix E.1 for model introduction. The principle of model selection is to ensure strong conversational abilities in Chinese, given that our benchmark involves variable-turn dialogue.. Therefore, models with poor Chinese proficiency, such as Llama2 (Touvron et al., 2023), or models with weak dialogue capabilities, such as Mixtral (Jiang et al., 2024), are not selected. The temperature for the user-simulator (GPT-4) is set to 0 to ensure consistency in generating the first query. For closed-sourced models like the GPT series, we leveraged the official API for inference. For open-sourced models, we deployed them locally, utilizing two A100 80GB PCIe GPUs for inference. In addition, an another empirical study in Appendix C.2 demonstrates the feasibility of GPT-4 as user-simulator.

4.2 Main Results

Open-source General LLMs: Table 3 shows the benchmarking results. The emergence of numerous open-source general LLMs presents a compelling alternative to closed-source models like ChatGPT and GPT-4, particularly within user-centric legal scenarios. Especially notable are the latest smaller models, such as Qwen-2.5-7B-Instruct, which have achieved performance comparable to OpenAI o1-preview. The results suggest that these general-purpose models, even without specific legal training, can effectively address user-centric legal issues and find practical applications. The availability of these open-source alternatives empowers legal professionals with greater flexibility and

²More details are shown in Appendix D.

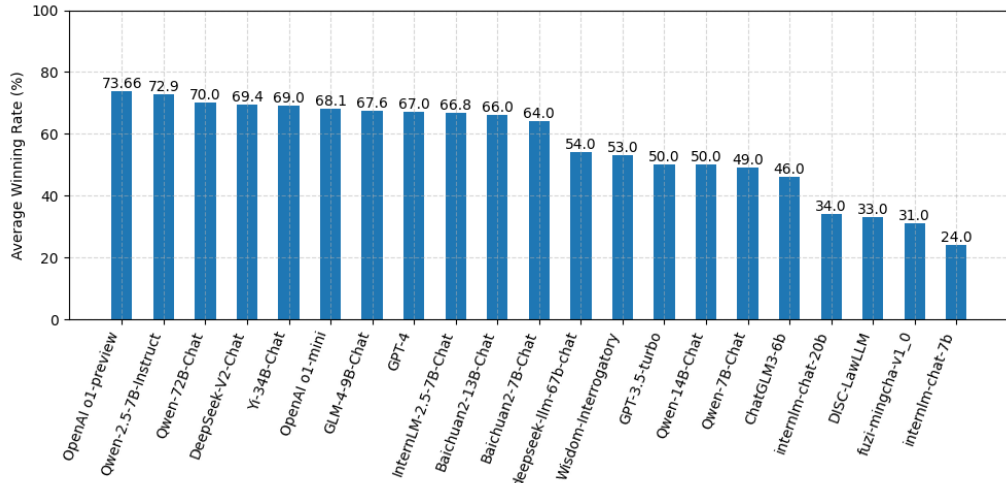


Figure 5: Average winning rate of 15 LLMs evaluated on UCL-bench. The baseline model ChatGPT defaults to a 50% win rate.

	Article Recite	Legal Issue Consult	Judicial Interpret	Legal Concept Und.	Public Opinion Sum.	Judicial Sum.	Prison Term Pre.	Charge Predict	Accuse Predict	Article Predict	Dispute Focus Id.	Real Case Consult	Similar Case Retrieve	Legal Case Analysis	Judicial Exam Score	Judgment Doc. Pt.	Legal Essay Write	Cross-Exam. Gen.	Litigation Strategy Gen.	Judgment Doc. Gen.	Prosecutor Doc. Gen.	Lawyer Doc. Gen.	AVG
Qwen 7B	40.0	52.5	52.5	40.0	55.0	25.0	53.1	32.5	57.5	60.0	15.0	29.4	42.1	30.0	30.0	22.5	10.0	30.0	25.0	61.1	81.6	50.0	40.7
Qwen 14B	75.0	50.0	62.5	47.5	30.0	40.0	50.0	30.0	55.2	67.5	20.0	34.3	32.5	50.0	45.0	47.5	15.0	60.0	25.0	73.7	92.5	38.2	47.3
Qwen 72B	60.0	60.0	72.5	55.0	57.5	72.5	47.1	40.0	72.5	81.6	32.5	58.9	42.1	67.5	50.0	47.5	30.0	55.0	37.5	85.0	68.4	55.9	56.7
Qwen-2.5 7B	85.8	69.7	70.3	70.0	68.9	70.7	69.0	80.6	82.5	81.2	75.4	67.6	67.9	70.6	71.1	72.1	69.8	78.0	69.1	75.7	70.1	67.9	72.9
Yi 34B	72.5	77.5	57.9	52.5	47.5	72.5	43.8	60.0	76.3	68.4	50.0	41.2	37.5	71.1	52.5	47.5	65.0	65.0	52.5	70.0	77.5	47.1	59.3
Baichuan2 7B	67.5	57.5	47.5	47.5	52.5	65.0	67.6	22.5	80.0	72.5	37.5	31.3	35.0	45.0	47.5	30.0	30.0	70.0	50.0	72.5	92.5	61.8	53.8
Baichuan2 13B	75.0	67.5	75.0	36.8	50.0	55.0	38.2	52.5	77.5	85.0	57.5	44.1	25.0	55.0	42.5	47.5	55.0	60.0	57.5	73.7	81.6	58.9	57.7
Deepseek 67B	50.0	60.0	60.0	57.5	57.5	75.0	55.9	47.5	67.5	50.0	50.0	47.1	37.5	50.0	50.0	60.0	20.0	35.0	45.0	42.1	60.5	56.3	51.6
DeepSeek-V2-Chat	78.3	66.9	67.0	67.5	65.5	67.5	66.2	74.0	77.5	76.2	71.4	65.7	65.2	67.7	69.2	70.7	67.1	72.7	66.1	70.5	66.6	65.8	69.3
ChatGLM3 6B	62.5	42.5	52.5	73.7	55.0	75.0	50.0	55.0	45.0	62.5	45.0	59.4	52.5	50.0	49.0	30.0	44.4	35.0	52.7	52.5	67.5	67.6	53.2
GLM-4-9B-Chat	74.1	66.7	67.0	66.2	64.3	66.8	65.8	68.7	75.0	77.5	66.3	64.1	63.7	66.7	66.4	67.3	66.3	69.0	66.8	67.9	66.1	63.7	67.6
Internlm 7B	52.5	42.5	40.0	45.0	13.1	5.0	38.2	45.0	65.0	42.5	17.5	47.1	40.0	36.9	25.0	47.4	15.0	15.0	36.8	22.2	75.0	20.6	35.8
Internlm 20B	35.0	41.7	50.0	40.0	27.5	45.0	47.1	42.5	55.0	42.5	37.5	53.0	31.6	32.5	44.4	30.0	10.0	50.0	42.5	26.3	35.0	38.2	38.9
InternLM-2.5 7B	75.0	64.3	65.4	65.4	63.9	65.4	63.7	71.7	75.0	78.9	66.1	63.7	63.7	63.7	64.6	65.9	64.8	67.3	64.2	66.4	64.5	64.1	66.8
GPT-4	62.5	50.0	62.5	57.5	67.5	81.6	61.8	27.5	65.0	72.5	42.5	61.8	45.0	57.5	52.6	57.5	50.0	60.0	37.5	37.5	65.0	58.8	56.1
OpenAI o1-mini	77.5	71.7	69.2	68.4	67.5	66.6	66.6	63.3	65.0	65.0	66.6	68.2	68.2	67.4	69.1	70.0	69.2	69.2	68.4	67.4	68.7	68.7	
OpenAI o1-preview	82.5	83.9	82.5	78.4	76.8	75.0	72.2	71.8	72.2	71.8	73.4	73.7	75.0	74.4	74.4	75.0	75.0	74.6	73.4	72.5	71.6	71.6	73.2
Wisdom. 7B	65.0	68.4	47.5	42.5	50.0	26.3	44.1	62.5	65.0	77.5	37.5	76.5	35.0	47.5	50.0	55.0	20.0	40.0	47.5	55.0	72.5	35.3	50.9
DISC-Law 13B	52.5	52.6	52.5	22.5	22.5	25.0	41.2	27.5	55.0	45.0	17.5	41.2	18.4	57.5	35.0	39.5	30.0	40.0	47.4	17.5	62.5	23.5	37.4
Fuzi-Mingcha 7B	75.0	20.0	27.5	30.0	12.5	36.8	43.3	50.0	65.0	60.0	7.5	26.5	7.5	39.5	15.0	12.5	10.0	45.0	30.0	60.0	60.0	14.7	34.0

Table 3: Evaluation results of UCL-bench. The values in the table represent the win rates (%) of LLM against ChatGPT. Ties were factored in by treating each tie as a 50% win to ensure stable results.

choice, expanding their options beyond solely relying on closed-source models like ChatGPT or GPT-4. There is a moderate positive correlation between the general and legal-specific performance, see Appendix E.1 for detail.

Legal LLMs: Despite being pre-trained or fine-tuned on legal data, existing legal LLMs have not yet surpassed ChatGPT and these open-source general LLMs in addressing legal issues within user-centric scenarios, their overall scores generally fall below ChatGPT, let alone open-source general LLMs. Specifically, only Wisdom-Interrogatory exhibits slightly better performance than ChatGPT. It is based on the Baichuan 7B base model and was pre-trained on 40GB of legal data, including legal documents, judicial cases, and legal Q&A data. Consequently, it significantly outperforms ChatGPT in tasks that emphasize legal knowledge and format, such as legal provision memory

(65%), crime/provision prediction (62.5%, 77.5%), and judgment/prosecutorial document generation (55.0%, 72.5%). However, it struggles in tasks requiring creative thinking, such as legal essay writing (20.0%), dispute focus identification (37.5%), and lawyer document generation (35.3%). Other legal LLMs, such as DISC-LawLLM, despite being fine-tuned on the Baichuan 13B base model with traditional NLP tasks like information extraction, text summarization, case classification, and legal open-domain question answering data, as well as judicial exam data, have not shown a clear advantage over ChatGPT in related tasks. This suggests that current legal LLMs may not be meeting expectations in user-centric scenarios and may not be readily replacing ChatGPT in most cases. This discrepancy could be attributed to the fact that these legal LLMs were developed using base models that were not as extensively trained on diverse data as

today’s open-source general LLMs. Furthermore, it may also indicate these pre-training and fine-tuning methods, while enhancing legal knowledge, might not necessarily improve the LLMs’ ability to effectively solve user problems.

Fine-grained discussion: From the results, we can find that while legal LLMs (DISC-Law13B, Fuzi-Mingcha7B) exhibit relatively high win rates in memorization tasks (Article Recite; 52.5% and 75%, respectively), their performance significantly lags behind general LLMs in higher-order cognitive tasks (understanding and application). Analysis of individual cases indicates that legal LLMs excel at recalling information from legal texts due to their specialized training data. However, general LLMs demonstrate superior multi-task capabilities, exhibiting strong reasoning abilities even in the legal domain when provided with sufficient information. Furthermore, although legal LLMs often provide correct final answers, they frequently omit key points identified in the answer guidance, indicating a lack of focus on crucial legal details. Their tendency to overlook information within the questions also necessitates additional conversational turns to clarify missing information, leading to lower evaluator scores.

Dialogue Turns: Our benchmarking results reveals that stronger models tend to resolve users’ legal needs through shorter dialogues. As shown in Figure 6, the average number of dialogue turns exhibits a decreasing trend with the increase in the win rate, with a correlation coefficient of -0.68 and a p-value of 0.005, indicating a fairly strong negative correlation between them. While more turns in dialogue can gather more information, this is also related to the LLM’s understanding and expression capabilities. The current results seem to indicate that LLMs with more turns have poorer performance, suggesting that these additional turns are likely due to the LLM’s inability to resolve the issue. An in-depth observation of the dialogues also support this judgement. For example, Internlm-chat-7b tends to provide overly brief responses, like only mentions "law xx" without providing specific articles (see Table 36); DISC-LawLLM refuse to give helpful answers in some circumstances (see Table 37). To the best of our knowledge, we are the first to conduct LLM evaluation in legal scenarios through simulated dialogues and explain the strong correlation between the number of turns and the LLM’s legal capabilities.

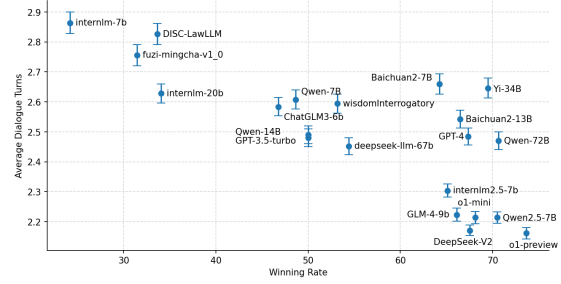


Figure 6: The relationship between the average number of dialogue turns and the winning rate among 15 LLMs.

4.3 Effectiveness of LLM-as-the-judge

We evaluated the effectiveness of GPT-4 judge from both contents and results it provides, through expert annotation methods. A fine-grained analysis can be found in Appendix E.2. The results show that GPT-4 judge mentioned 95.9% of the answer guidance points and human experts believed that GPT-4 correctly determined whether the LLM met the points in 80.6% of cases. The agreement rate between GPT-4 and human experts is 81.1%, which is very close to the agreement among humans (80.1%). These results indicate that GPT-4 judge has the ability to provide effective and reliable evaluation results.³

5 Conclusion

We introduced UCL-bench, a user-centric legal benchmark within the Chinese legal system, comprising 22 tasks that collectively cover a wide range of use cases in the legal field. User surveys were conducted to help build legal datasets that reflect practical needs. A user-simulator was utilized for multi-turn interactions with LLMs. Using answer guidance as a reference, GPT-4 was employed to assess the performance of the LLMs. Experimental results demonstrate that state-of-the-art open-sourced LLMs excel on UCL-Bench, proving their applicability under the Chinese legal framework. Despite being pre-trained and fine-tuned with legal data, legal LLMs fell short of ChatGPT, highlighting the need for training strategies more attuned to user needs. Additionally, models that deliver concise and accurate responses typically resolve legal issues in fewer dialogue turns, which is essential for achieving high performance in user-centric scenario. We hope our benchmark can serve as a

³We also conducted additional experiments in Appendix E.2 to investigate the feasibility of replacing GPT-4 with alternative models for evaluation.

reference for the capabilities of LLMs in addressing user-centric legal issues and provide insights on enhancing the utility of legal LLMs. In the future, we aim to expand UCL-bench to other languages, incorporating a broader range of legal scenarios across different legal frameworks.

6 Limitations

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

- While we present extensive further experiments with our simulator and evaluator in Appendix, we are currently unable to fully replace the simulator and evaluator with human participants due to the complexity and cost associated with multi-turn dialogues.
- We use ChatGPT as a baseline in our testing, comparing other LLMs to its performance. As ChatGPT's performance has become increasingly insufficient compared to newer LLMs, we will attempt to use updated models for comparison in the future.
- Due to the difficulty of user survey, data collection, and construction, we have currently only built a Chinese dataset. We plan to build datasets for more languages in the future.
- Our evaluation primarily relies on win rate, which is currently the main metric for assessing LLMs. However, compared to other results presented in our Appendix, we find that win rate may not fully capture the effectiveness of LLMs. Therefore, we are committed to exploring additional and more comprehensive evaluation metrics in the future.

7 Ethics Statement

Recognizing the sensitive nature of the legal domain, we have meticulously reviewed all data included in this benchmark. The open-source datasets employed are accompanied by appropriate licenses. To ensure privacy and ethical considerations, we have masked personal sensitive information, such as names, contact details, and identification numbers. Furthermore, legal experts have conducted thorough ethical evaluations of the benchmark, ensuring its responsible use.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. [Yi: Open foundation models by 01.ai](#).
- Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Baichuan. 2023. [Baichuan 2: Open large-scale language models](#). *arXiv preprint arXiv:2309.10305*.
- Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiusi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.
- Kelley Burton. 2017. "think like a lawyer" using a legal reasoning grid and criterion-referenced assessment rubric on irac (issue, rule, application, conclusion). *Journal of Learning Design*, 10(2):57–68.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaying Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai

- Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingdong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. 2024. [Internlm2 technical report](#).
- Ilias Chalkidis, Abhik Jana, Dirk Hartung, M Bommaritto, Ion Androutsopoulos, D Martin Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. arxiv.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.
- Inyoung Cheong, King Xia, KJ Feng, Quan Ze Chen, and Amy X Zhang. 2024. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. *arXiv preprint arXiv:2402.01864*.
- Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. [Chatlaw: Open-source legal large language model with integrated external knowledge bases](#).
- Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2023. Laiw: A chinese legal large language models benchmark (a technical report). *arXiv preprint arXiv:2310.05620*.
- DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. 2024. [Deepseek llm: Scaling open-source language models with longtermism](#).
- Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echevoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.
- 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.
- 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*.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.
- 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 Rasmov-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*.
- Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.
- Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *ArXiv*, abs/2305.15062.

- Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. Il-tur: Benchmark for indian legal text understanding and reasoning. *arXiv preprint arXiv:2407.05399*.
- Charles Kakilla. 2021. Strengths and weaknesses of semi-structured interviews in qualitative research: A critical essay.
- David R Krathwohl. 2002. A revision of bloom’s taxonomy: An overview. *Theory into practice*, 41(4):212–218.
- Aditya Kuppaa, Nikon Rasumov-Rahe, and Marc Voses. 2023. Chain of reference prompting helps llm to think like a lawyer. In *Generative AI+ Law Workshop*.
- Frans L Leeuw and Hans Schmeets. 2016. *Empirical legal research: A guidance book for lawyers, legislators and regulators*. Edward Elgar Publishing.
- 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.
- Bulou Liu, Yiran Hu, Qingyao Ai, Yiqun Liu, Yueyue Wu, Chenliang Li, and Weixing Shen. 2023a. Leveraging event schema to ask clarifying questions for conversational legal case retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1513–1522.
- Bulou Liu, Yiran Hu, Yueyue Wu, Yiqun Liu, Fan Zhang, Chenliang Li, Min Zhang, Shaoping Ma, and Weixing Shen. 2023b. Investigating conversational agent action in legal case retrieval. In *European Conference on Information Retrieval*, pages 622–635. Springer.
- Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.
- Michael D Murray. 2023. Artificial intelligence for academic support in law schools and universities. Available at SSRN 4564227.
- 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. *arXiv preprint arXiv:2301.13126*.
- Konstantinos I Roumeliotis and Nikolaos D Tselikas. 2023. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192.
- Nancy L Schultz. 1992. How do lawyers really think? *J. Legal Educ.*, 42:57.
- 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*.
- Weiwei Sun, Shuyu Guo, Shuo Zhang, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2023a. Metaphorical user simulators for evaluating task-oriented dialogue systems. *ACM Transactions on Information Systems*, 42(1):1–29.
- Yuchong Sun, Che Liu, Jinwen Huang, Ruihua Song, Fuzheng Zhang, Di Zhang, Zhongyuan Wang, and Kun Gai. 2023b. Parrot: Enhancing multi-turn chat models by learning to ask questions. *arXiv preprint arXiv:2310.07301*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.
- 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>.
- Wu Yiquan, Liu Yuhang, Liu Yifei, Li Ang, Zhou Siying, and Kuang Kun. *wisdominterrogatory*. Available at GitHub.
- 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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Details of User Survey

A.1 Semi-structured Interview

The purpose of the semi-structured interview is to identify the legal scenarios and their use cases for LLMs. A total of eleven legal experts participated in the interviews, including three law professors, two developers of legal LLMs, two lawyers, two judges, and two procurators. Table 4 shows the demographic background of the participants, with 9 males and 2 females. The majority of the participants fall within the age range of 30-50 and possess over 10 years of work experience. All participants hold a bachelor's degree or higher.

The first author used an interactive demonstration method to introduce LLM to participants at the beginning of the interview. This involved giving the participants the opportunity to engage with ChatGPT by inputting prompts or questions and observing the responses generated in real-time. The interviews were conducted in Mandarin and lasted 50-60 minutes each. Every participant was paid 200 yuan after the interview. Below is the interview protocol for the participants.

1. Background Information and Self-introduction

- Purpose of the Interview: We aim to understand the main users in the legal field and respective use cases for LLMs.
- Self-Introduction: Please introduce yourself, including your profession, workplace, and years of experience in the legal field.
- Main Job Responsibilities: What legal-related tasks are you primarily responsible for in your daily work?

2. User Roles

- Essential Roles: What are the indispensable roles in the Chinese legal system?
- Coverage of Legal Scenarios: Do these roles cover most legal scenarios? If not, what additional roles are needed?

- Familiar Roles: Which of these roles do you interact with and are familiar with in your daily work?

3. Legal Issues in Daily work

- Daily Workflow: What is the daily workflow in your profession? What tasks need to be completed?
- Legal Issues: What issues arise in your daily work? For example, tasks that are particularly difficult or time-consuming.
- Legal issues of Other Roles: What are the workflows and tasks of other legal professions? What types of legal issues do they encounter?

4. Experience with AI Technology

- Experience with Technology: Do you have experience using legal artificial intelligence tools? If so, which tools do you use?
- Currently Used Tools: What legal technology tools are you currently using? What roles do they play in your work?
- Advantages and Disadvantages of Tools: What are the strengths and weaknesses of these tools? Are there areas that need improvement?

5. Needs for Using LLMs

- Understanding of Large Models: How familiar are you with LLMs?
- Usage Experience: Have you used LLMs to solve any problems in your daily work? If so, how effective were they?
- Potential Use Cases for LLMs: LLMs can aid in legal document drafting, contract review, legal research, case prediction, etc. In your opinion, what are the use cases for LLMs in the legal field?

6. Future Outlook

- Expectations for Technological Development: What are your expectations for the development of legal technology in the coming years?
- Suggestions and Opinions: Do you have any suggestions or opinions on how we can better develop and apply large models to meet the needs of the legal field?

	Occupation	Gender	Age Group (years)	Working Experience (years)	Educational Background
T1	Professor	Male	30-40	5-10	Phd
T2	Professor	Male	40-50	More than 10	Phd
T3	Professor	Female	40-50	More than 10	Phd
T4	Algorithm Engineer	Male	30-40	5-10	Phd
T5	Algorithm Engineer	Male	20-30	0-5	Master
T6	Lawyer	Male	20-30	5-10	Bachelor
T7	Lawyer	Male	30-40	More than 10	Master
T8	Judge	Male	50-60	More than 10	Bachelor
T9	Judge	Male	30-40	More than 10	Master
T10	Procurator	Female	30-40	More than 10	Master
T11	Procurator	Male	40-50	More than 10	Bachelor

Table 4: The demographic background of the participants.

A.2 Questionnaire Design

We have designed an online questionnaire to gather insights into the tasks and scenarios where users are interested or believe that LLMs can provide assistance. The survey was conducted from January 10, 2024, to January 24, 2024, lasting for two weeks. Before the questionnaire survey, we provided all participants with GPT-4 accounts to assist them in their work. Therefore, it can be assumed that they all have some understanding of the capabilities and characteristics of LLMs. A total of 346 responses were collected. We excluded responses where the completion time was less than five minutes as invalid, retaining 282 valid responses. Among the retained responses, 56 were from the general public, 62 from colleges, 64 from law offices, 52 from courts, and 48 from procuratorates. Below is the structure of the questionnaire.

Title: Survey on the Use of LLMs in Legal Field

Introduction: Thank you for participating in our survey. The purpose of this survey is to understand the use of LLMs in the legal field. Your answers will provide valuable data for our research and help us better design and improve relevant technologies. This questionnaire will take approximately 5-10 minutes to complete and your responses will be kept strictly confidential.

Main Contents:

- Which of the following five scenarios best describes you? **(Single choice)**
 - General public
 - Law students/professors
 - Law office
 - Court
 - Procuratorate
- (Exclusively For General Public)** In your daily life, if you have legal needs, LLM can offer a variety of legal services, significantly enhancing the accessibility of legal resources. Among the following scenarios, which ones

are you most interested in or believe that LLM could provide support to the general public?

(Multiple Choice)

- Ask general legal issues, like proceedings, rights and obligations
- Seek legal advice or solutions for specific cases
- Explain legal terms and concepts
- Assist in completing and customizing legal documents
- Seek opinion on widely-publicized legal case

- (Exclusively For College)** In your daily studies and work, LLM can serve as your study and research assistant, offering support in areas such as legal research, case study analysis, and preparation for examinations. Among the following scenarios, which ones are you most interested in or believe that LLM could provide support to law students or professors?

(Multiple Choice)

- Help with legal essay writing
- Analyze specific legal theories, principles, or concepts
- Analysis of historical or recent legal case
- Suggest for improvement of the current legislation
- Grading legal exam questions
- Find scholarly articles, case rulings on specific legal issues

- (Exclusively For Law office)** In your daily work, LLM can become a valuable ally, empowering your work across a wide range of fields with its capabilities in legal research, document drafting, and case analysis. Among the following scenarios, which ones are you most interested in or believe that LLM could provide support to law office? **(Multiple Choice)**

- Search for historical precedents similar to

the client's case.

B: Seek the relevant legal provisions of a particular case.

C: Assist in drafting legal documents such as pleadings, contracts.

D: Provide legal information and advice to clients.

E: Query the specific content of the legal provisions.

F: Assist in proofreading legal documents for errors.

5. **(Exclusively For Courts)** In your daily work, LLM can assist in managing case files, simplifying legal research, and helping to draft preliminary judgments or opinions, thereby enhancing the efficiency of court operations. Among the following scenarios, which ones are you most interested in or believe that LLM could provide support to courts? **(Multiple Choice)**

A: Assist judges in deciding charges of defendants

B: Assist judges in deciding sentences of defendants

C: Help draft judgments and other legal documents

D: Summarizing cases for filing

E: Answer legal questions from the public

F: Inquire about the litigation procedure for a case

6. **(Exclusively For Procuratorate)** In your daily work, LLM can help in analyzing case files, preparing legal documents, and offering insights into legal precedents, thus streamlining the prosecutorial process. Among the following scenarios, which ones are you most interested in or believe that LLM could provide support to Procuratorate? **(Multiple Choice)**

A: Precise interpretation of specific legal provisions

B: Formulate effective public prosecution strategy

C: Assist in preparing indictments and other legal documents

D: Assist in determining accusations of defendants

E: Analyze the public reactions to legal issues

7. Beyond the scenarios mentioned above, are

there any other use cases you are particularly interested in? **(Open-ended question)**

Use Cases: _____

A.3 Use Cases in the Legal Domain

We collected a total of 33 use cases through semi-structured interviews and questionnaires. By calculating the proportion these use cases practitioners were interested in and retaining those with a proportion greater than 10%, we ultimately identified 22 use cases as potential high-frequency applications of LLMs in the legal field. The final determined use cases are shown in the Table 5.

B Dataset Construction

B.1 Details of Task Verification

Multiple rounds of questionnaire surveys were conducted to validate the effectiveness of the task design. The survey took place from February 1, 2024, to February 14, 2024, lasting for two weeks. The participants were the same group as in our previous user survey.

Participants were first asked to rate the clarity of task descriptions within their respective scenarios, the relevance to their use cases, the degree to which the tasks met their actual needs, using a 5-point Likert scale (1-5), and an open-ended question solicited their suggestions for task improvements. We identified tasks with an average score below 4 in any dimension and had legal experts modify these tasks based on user feedback. These modified tasks were then subjected to further validation through subsequent questionnaires until all tasks meet the requirements.

The survey was conducted three rounds in total. Ultimately, the average scores for clarity, relevance to use cases, and degree of meeting actual needs for each task all exceeded 4, empirically validating the effectiveness of our task design. The link between use cases and tasks is shown in Table 5; the verification scores are shown in Table 6. Below is the structure of the task validation survey.

Title: Initial task validation survey

Introduction: Thank you for taking part in our survey. Based on the results of the previous questionnaire survey, we identified 22 high-frequency use cases and asked legal experts to design corresponding legal tasks according to these use cases. Now we would like to ask you to evaluate the task to verify the effectiveness of the task design. The

Use Case	Corresponding Task
Ask general legal issues like proceedings and rights	Legal Issue Consult
Seek legal advice or solutions for specific cases	Real Case Consult
Help with legal paper writing	Legal Essay Writing
Analyze specific legal theories, principles, or concepts	Legal Concept Understand
Analysis of historical or recent legal cases	Legal Case Analyze
Suggest for improvement of the current legislation.	Litigation Strategy Gen.
Grading legal exam questions	Judicial Exam Score
Search for historical precedents similar to the client's case	Similar Case Retrieval
Seek the relevant legal provisions of a particular case	Article Prediction
Query the specific content of the legal provisions	Article Recitation
Assist in drafting legal documents such as pleadings , contracts	Lawyer Document Gen.
Assist in proofreading legal documents for errors	Legal Doc. Proofread
Guidance on how to legally collect and effectively evaluate evidence	Cross-Examination Gen.
Assist judges in deciding charges of defendants	Charge Prediction
Help draft judgments and other legal documents	Judgment Doc Gen.
Assist judges in deciding sentences of defendants	Prison Term Prediction
Identify the key issues in dispute of the case	Dispute Focus Identify
Summarizing cases for filing	Judicial Summary
Precise interpretation of specific legal provisions	Judicial Interpret
Assist in preparing indictments and other legal documents	Procuratorial Document Gen.
Analyze the public reactions to legal issues	Public Opinion Summary
Assist prosecutors in determining accusations of defendants	Accusation Prediction

Table 5: Final Determined Use Cases Corresponding Tasks.

questionnaire will last 5-10 minutes and your responses will be kept strictly confidential.

(Basic Information)

- Which of the following five scenarios best describes you? (Single choice)
 - General public
 - Law students/professors
 - Law office
 - Court
 - Procuratorate

(Task Evaluation for a specific task.)

Use case: {use case}

Task name: {task name}

Task description: {task description}

- How clear is the task description?
 - Very unclear
 - Unclear
 - Neutral
 - Clear
 - Very clear
- How relevant is the task to the use case?
 - Not relevant
 - Slightly relevant
 - Somewhat relevant
 - Relevant
 - Highly relevant
- How well does the task meet your actual needs?

- Not at all
- Slightly
- Moderately
- Well
- Very well

- Do you have any suggestions for improving this task?(Open-ended question)
Suggestions: _____

B.2 Bloom Taxonomy

Bloom's Taxonomy is a multi-tiered model used to classify educational learning objectives into levels of complexity and specificity (Kratwohl, 2002). It was developed by Benjamin Bloom in 1956 and later revised by Lorin Anderson in 2001. The taxonomy is divided into six levels, arranged from the simplest to the most complex: Remember, Understand, Apply, Analyze, Evaluate, and Create. These levels represent a progression from basic knowledge acquisition to higher-order thinking skills, encouraging educators to design learning experiences that promote deep understanding and critical thinking. In the context of developing law benchmarks, this structured approach aligns well with the hierarchical nature of legal reasoning and understanding, thereby helping create a diverse set of tasks that can effectively assess different levels of cognitive abilities and skills of the LLMs, ranging from basic knowledge recall to complex problem-solving and critical thinking, as shown in Figure 7. It is

Corresponding Task	Clarity	Relevance	Practicality
Legal Issue Consult	4.7	4.2	4.5
Real Case Consult	4.6	4.5	4.3
Legal Essay Writing	4.5	4.1	4.2
Legal Concept Understand	4.6	4.3	4.4
Legal Case Analyze	4.7	4.2	4.2
Litigation Strategy Gen.	4.2	4.6	4.7
Judicial Exam Score	4.6	4.7	4.6
Similar Case Retrieval	4.5	4.1	4.2
Article Prediction	4.2	4.4	4.4
Article Recitation	4.7	4.2	4.1
Lawyer Document Gen.	4.6	4.7	4.7
Legal Doc. Proofread	4.2	4.3	4.2
Cross-Examination Gen.	4.3	4.2	4.2
Charge Prediction	4.1	4.2	4.4
Judgment Doc Gen.	4.4	4.1	4.3
Prison Term Prediction	4.7	4.2	4.0
Dispute Focus Identify	4.2	4.5	4.1
Judicial Summary	4.6	4.2	4.0
Judicial Interpret	4.5	4.4	4.5
Procuratorial Document Gen.	4.6	4.1	4.0
Public Opinion Summary	4.6	4.2	4.1
Accusation Prediction	4.4	4.2	4.1

Table 6: Task Verification Scores. Clarity indicates the precision and comprehensibility of the task description, Relevance reflects the alignment between the task and the intended use case, and Practicality assesses how well the task meets real-world needs. Scores range from 1 to 5, with higher scores indicating higher effectiveness.

worth noting that a single task may examine the LLM’s capabilities across multiple cognitive levels. In our approach, we categorize each task based on its most dominant cognitive level.

B.3 Task Description

Based on the use cases derived from user surveys, we have constructed a benchmark dataset. A total of 22 tasks were gathered categorized them according to the Bloom’s cognitive taxonomy, as illustrated in Table 8. Below are specific descriptions of each task.

Data Source	License	Proportion
flk.npc.gov.cn	public	9.1%
imlaw.cn	public	4.5%
moj.gov.cn	public	18.2%
wenshu.court.gov.cn	public	27.3%
12309.gov.cn	public	9.1%
12348.gov.cn	public	4.5%
CAIL2021	MIT License	9.1%
CAIL2022	MIT License	9.1%
LeCaRD	MIT License	4.5%

Table 7: Licenses of our data source. Proportion represents the percentage of the dataset that originates from this data source.

Remember Article recitation(1-1) is the only one task in memory capability.

- **Article Recitation (1-1)** Recite the specific content of the legal provisions. *Information:*

An legal article number; Needs: Recite the article content. This task examines to which extent LLMs encode legal knowledge within their parameters. Data is sourced from legal provisions in flk.npc.gov.cn website. We collected samples in a dispersed manner, ensuring coverage of China’s major laws.

Understand There are 5 tasks in the understand category. A precise and in-depth understanding of legal articles and concepts is crucial for successfully completing downstream tasks.

- **Legal Issue Consultation (2-1)** Seek legal advice on general legal issues. *Information: None; Needs: Seek legal advice on a general legal question.* Legal issue consultation is crucial for helping people understand their rights and obligations, or legal procedure in Chinese legal framework. We collect general legal issues from legal articles in imlaw.cn to construct our dataset, covering a wide array of topics such as traffic violations, marriage, debt, and labor disputes.
- **Judicial Interpretation (2-2)** Interpret a legal term in a legal article. *Information: A specific legal article; Needs: Interpret a legal term in that legal article.* This task mirrors real-world scenarios where legal professionals interpret statutes for clients, aiding in understanding

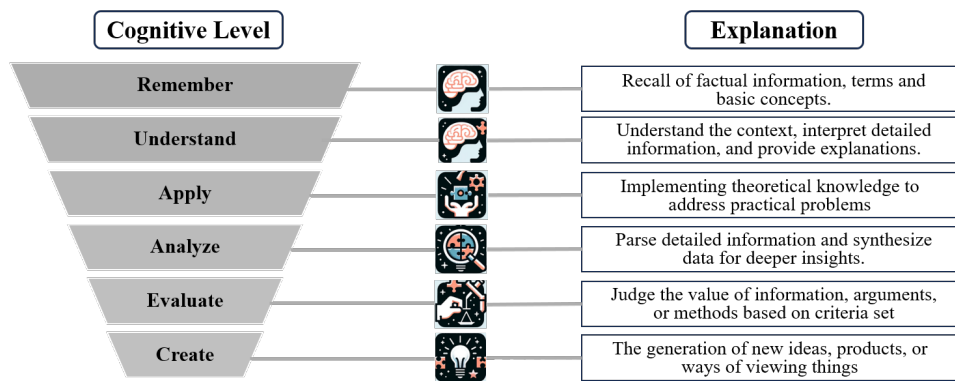


Figure 7: Bloom cognitive taxonomy.

Level	General Public	College	Law Office	Court	Procuratorate
Memory			Article Recitation		
Understand	Legal Issue Consult	Legal Concept Understand		Judicial Summary Judicial Interpret	Judicial Interpret Public Opinion Summary
Application				Prison Term Prediction Charge Prediction	Accusation Prediction
Analyze	Real Case Consult	Legal Case Analyze	Similar Case Retrieval Article Prediction	Dispute Focus Identify	
Evaluate		Judicial Exam Score	Legal Doc. Proofread		
Create		Legal Essay Writing Litigation Strategy Gen.	Cross-Examination Gen. Lawyer Document Gen.	Judgment Doc Gen.	Procuratorial Document Gen.

Table 8: The distribution of task sets at cognitive levels and scenarios.

how laws apply to specific situations. Data for this task is collected from judicial interpretations articles found within national legal databases.

- **Legal Concept Understanding (2-3)** Answer the conceptual questions of the judicial examination. *Information: A conceptual question from judicial examination; Needs: Get the answer.* This task reflects the practice legal students engage in preparation for judicial exams. The data for this task is sampled from the CAIL 2022 dataset, focusing on basic concepts within major laws such as constitutional law, civil law, commercial law, and criminal law.
- **Public Opinion Summary(2-4)** Generate summaries of articles related to legal opinion. *Information: An article related to legal public opinion; Needs: Generate a summary of the article* Summarizing legal public opinion articles is crucial for legal professionals and policymakers to grasp public sentiments on legal issues, aiding in responsive policy formulation. The data is directly sourced from the CAIL2022 dataset.
- **Judicial Summary (2-5)** Generate judicial summary of judgment document. *Informa-*

tion: A judgment document; Needs: Generate a judicial summary highlighting its key points and conclusions. Summarizing court judgment documents helps legal professionals and the public understand the essence of judicial decisions efficiently. The data for this task is directly sourced from the CAIL2021 dataset.

Apply In the legal domain, the apply level primarily involves applying laws and regulations to the facts of a case and deducing the consequent outcomes. We select three tasks for this category.

- **Prison Term Prediction (3-1)** Predict the sentence of the defendant according to the facts of the defendant’s crime. *Information: The criminal facts related to a defendant; Needs: Predict the likely prison sentence for the defendant.* This task is significant for judges as it aids in the sentencing process, contributing to more consistent and equitable legal proceedings. The data is sourced from the CAIL2021 dataset.
- **Charge Prediction (3-2)** Predict charges based on the facts of the defendant’s crime. *Information: The facts of the crime committed; Needs: Predict the charges.* It plays a critical role within the judicial framework, where judges determine the charges based on

the criminal facts presented. The data source consists of judgement documents available on China Judgments Online, from which we extract case facts and the related charges to compile this task.

- **Accusation Prediction (3-3)** Determine the specific charges and relevant legal provisions according to the relevant criminal facts of the criminal suspect. *Information: The criminal facts related to a suspect; Needs: Determine the specific accusations and relevant legal articles.* This task aids prosecutors in the precise drafting of indictment documents, ensuring the accuracy of legal charges. The data source includes indictment documents from the Supreme People's Procuratorate of China's website, from which verified facts and associated charges have been extracted.

Analyze At the "Analyze" level, tasks require a deep dive into legal documents and case specifics, extracting intricate legal relationships.

- **Article Prediction (4-1)** Determine the relevant legal provisions according to the facts of the crime. *Information: The facts of a case; Needs: Determine the relevant legal articles.* Identifying the correct legal articles quickly helps make the court's work more efficient and accurate. The data for this task is sampled from judgement documents in China Judgments Online, where we extract facts and relevant provisions of the case.
- **Dispute Focus Identification (4-2)** Determine the issue of the case based on the arguments of the plaintiff and the defendant. *Information: The arguments of the plaintiff and defendant; Needs: Identify the dispute focus of the case.* This task enhances legal clarity by pinpointing the core issues within disputes, directly impacting the efficiency of case resolutions. The dataset is derived from verdicts on China Judgments Online, extracting the arguments of both parties and the dispute focus as determined by the court.
- **Real Case Consultation (4-3)** Provide legal advice and insights based on real cases of clients. *Information: A real-world case; Needs: Provide consultation and insights tailored to the specifics of the case.* This task addresses the public's need for consulting on

actual cases, emphasizing the importance of analytical skills in handling unique legal scenarios. The dataset is sourced from publicly available consultation data on the China Legal Services website, where users ask legal questions and professional lawyers answer them online.

- **Similar Case Retrieval (4-4)** Determine which case is more relevant to the original case. *Information: An original case and two similar cases; Needs: determine which of the similar cases is more closely related to the original case.* This task is crucial for judges and lawyers who search for similar cases as references for their work, underlining the significance of comparative analysis in legal practice. The data is sampled directly from the LeCaRD dataset.
- **Legal Case Analysis (4-5)** Legal analysis of the case questions in the judicial examination. *Information: A case scenario from judicial examination; Needs: Perform a comprehensive legal analysis.* This task directly engages with case analysis questions of judicial examinations, assisting students in mastering the application of law. The data is obtained from the case analyze questions of past judicial examinations.

Evaluate At the "Evaluate" level, tasks are centered around assessing the quality and relevance of legal work. This stage demands the accuracy of legal writings and the alignment of judicial work with legal standards.

- **Judicial Examination Scoring (5-1)** Grading the subjective answers in judicial exam. *Information: A judicial examination essay question and its answer; Needs: Score the answer.* This task automates the scoring of subjective answers in judicial examinations, facilitating a rapid self-assessment for candidates regarding their legal knowledge. We engaged five law master's students to each complete ten case analysis questions from judicial exams as a data source.
- **Judgment Document Proofreading (5-2)** Proofread fragments of legal documents. *Information: A segment of a judgment document; Needs: Proofread it for accuracy and consistency.* This task reflects the routine

of court clerks proofreading judgment documents, crucial for enhancing the judiciary's authority by ensuring documents are error-free and coherent. The data is created by manually modifying actual judgment documents to introduce various errors, including punctuation, word order, logical inconsistencies, and typos.

Create At the "Create" level, tasks involve generating new legal ideas, strategies, or documents, requiring a high degree of innovation and critical thinking. We select six tasks in this level.

- **Legal Essay Writing (6-1)** Complete essay questions for the judicial exam. *Information: An essay question from a judicial examination; Needs: write a legal essay addressing the specified topic.* This task reflects the challenge faced by law students when tackling essay questions in judicial exams. The data for this task is sourced from essay questions of past Chinese judicial examinations.
- **Cross-Examination Strategy Generation (6-2)** Based on the facts and evidence related to the case, put forward cross-examination opinions on the evidence. *Information: Facts of the case and available evidence; Needs: Develop a strategy for cross-examining the evidence presented.* This task is designed to support lawyers in crafting specific objections to evidence, directly impacting the strength and direction of legal arguments in trial settings. The data is extracted from verdict documents on China Judgments Online, specifically focusing on the sections where the plaintiff and defendant present evidence and raise objections.
- **Litigation Strategy Generation (6-3)** Propose a legislative strategy based on the background information relevant to the legislation. *Information: Background related to legislation; Needs: Generate recommendations for litigation strategy.* This task supports legislative bodies in tailoring their legislative strategies based on the underlying legislative context, significantly influencing the formulation of laws and regulations. The data is sourced from legislative articles on the Ministry of Justice's official website, with extractions focusing on the legislative background and strategies discussed.

- **Judgment Document Generation (6-4)** Based on the facts of the case, generate the outcome part of the judgment. *Information: The facts of a case; Needs: Generate the adjudication results section of a judgment document.* This task facilitates the essential aspect of judicial writing by aiding in the structured formulation of verdicts. The data is extracted from the case facts sections of judgment documents.
- **Procuratorial Document Generation (6-5)** According to the criminal facts of the defendant, generate the conclusion of the indictment. *Information: The criminal facts of the defendant; Needs: Write the conclusion section of an indictment.* This task reflects the prosecutorial process of drafting the conclusion part of an indictment based on case facts, aiming to support writing efforts and enhance efficiency in legal documentation. The data is extracted from sections detailing the criminal facts and indictment charges in bill of complaints.
- **Lawyer Document Generation (6-6)** Based on the background of a legal event, generate the corresponding legal document. *Information: Legal event background; Needs: Generate a lawyer document.* The background of this task lies in the general public seeking to generate legal documents, including business contracts, complaints, and mediation agreements. The information and needs for the data are manually constructed by senior attorneys based on their professional experience.

C Details of User-LLM Interaction Simulation

C.1 Simulation Setup

GPT-4 is used as a user simulator to engage in multi-turn interactions with the model under evaluation, posing as a real user. At the outset, the user simulator obtains information-needs pair and presents the first question to the model. After the model provides an answer, the user simulator autonomously decides whether to continue the conversation or terminate it based on whether the user's needs are met.

During the conversation with the model, the user-simulator might affirm, explain, or correct the responses of the model during the dialogue process,

which are not commonly seen in real scenarios. Therefore, we restrict these behaviors by adding prompt, directing the user-simulator to focus on providing relevant information or asking questions. The prompt for user-simulator is shown in Figure 8 and translated in Figure 9.

In some tasks, "information" might be lengthy, making it challenging for the user simulator to replicate it accurately and in full. To address this issue, we present the "information" as a prompt to the model under evaluation in such long information tasks, allowing the user simulator to focus solely on posing questions based on user needs. In real-world scenarios, long text information often exists in the form of electronic records. Users can directly copy this information before asking questions, thus this approach does not compromise the realism of the simulation.

C.2 Feasibility of GPT-4 as User-Simulator

To evaluate the feasibility of GPT-4 as a user simulator, we enlisted a law master’s student to observe 1,100 dialogues and assess whether the questions posed by the user simulator are closely related to the needs. The student was compensated at a rate of 50 RMB per hour. The result shows that the user simulator almost always stays closely aligned with the user’s needs in the initial question and subsequently maintains relevance to the context in the following dialogues, providing effective information or posing follow-up questions related to the user’s needs. Failures occurred when the user simulator either answered questions itself or failed to provide relevant information. The statistical data is illustrated in Figure 9.

Counts	First Round	Subsequent Round
Total Number	1100	349
Aligned Number	1079	329
Aligned Ratio	98.1%	94.3%

Table 9: Statistics on User Simulator Performance.

D Evaluation Paradigm

The instruction template for prompting the GPT-4 evaluator is in Figure 10 and translated in Figure 11. To mitigate position bias, the positions of multi-turn dialogue data generated by the candidate and reference LLMs are assigned randomly. Moreover, to address the verbosity bias of GPT-4, i.e., the tendency to favor longer, more verbose responses, we instruct GPT-4 to base its judgments solely on

the consistency of AI assistant responses with the Answer Guidance.

E Details of Experiment

E.1 Models

We tested a total of 15 models, with 12 being general models and 3 legal-specialized models. The models’ information is shown in Table 10.

General LLMs

- **ChatGPT** A chatbot program developed by OpenAI, released on November 30, 2022. We use GPT3.5-turbo-1106 version for evaluation.
- **GPT-4** The fourth generation of natural language processing models released by OpenAI. We use GPT-4-0613 version for evaluation.
- **OpenAI o1-mini** A cost-efficient reasoning model released by OpenAI, o1-mini excels at STEM, especially math and coding—nearly matching the performance of OpenAI o1 on evaluation benchmarks such as AIME and Codeforces.
- **OpenAI o1-preview** o1-preview is OpenAI’s reasoning model for complex tasks. The model has 128K context and an October 2023 knowledge cutoff.
- **Baichuan2-Chat** Baichuan 2 series are open source LLMs launched by Baichuan Intelligence, which adopts high-quality corpus training of 2.6 trillion Tokens.
- **ChatGLM3** ChatGLM3 is a dialogue pre-training model jointly released by Zhipu AI and KEG Lab of Tsinghua University.
- **glm-4-9b-chat** GLM-4-9B is the open-source version of the latest generation of pre-trained models in the GLM-4 series launched by Zhipu AI.
- **deepseek-llm-chat** Deepseek LLM was trained from scratch on a vast dataset of 2 trillion tokens in both English and Chinese. It has 7B and 67B versions.
- **DeepSeek-V2-Chat** DeepSeek-V2.5 is an upgraded version that combines DeepSeek-V2-Chat and DeepSeek-Coder-V2-Instruct. The new model integrates the general and coding abilities of the two previous versions.

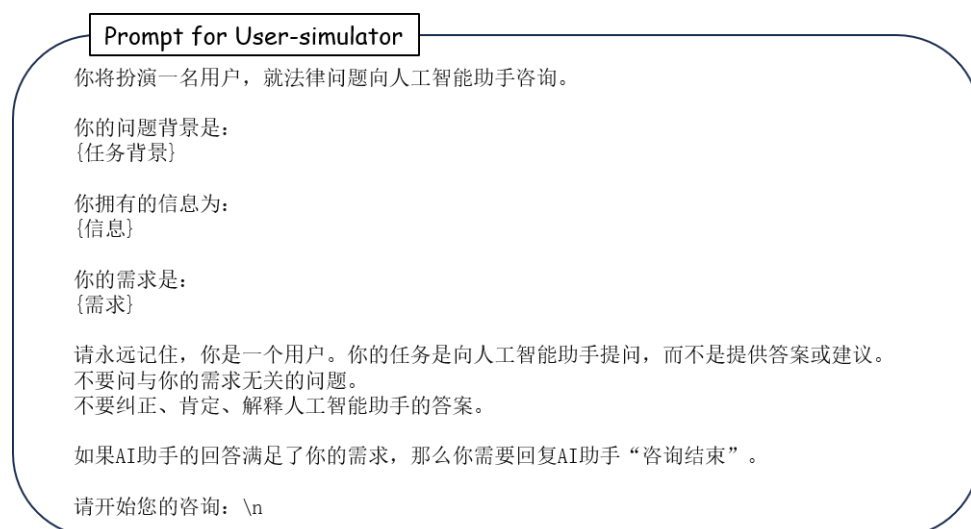


Figure 8: User-simulator prompt template.

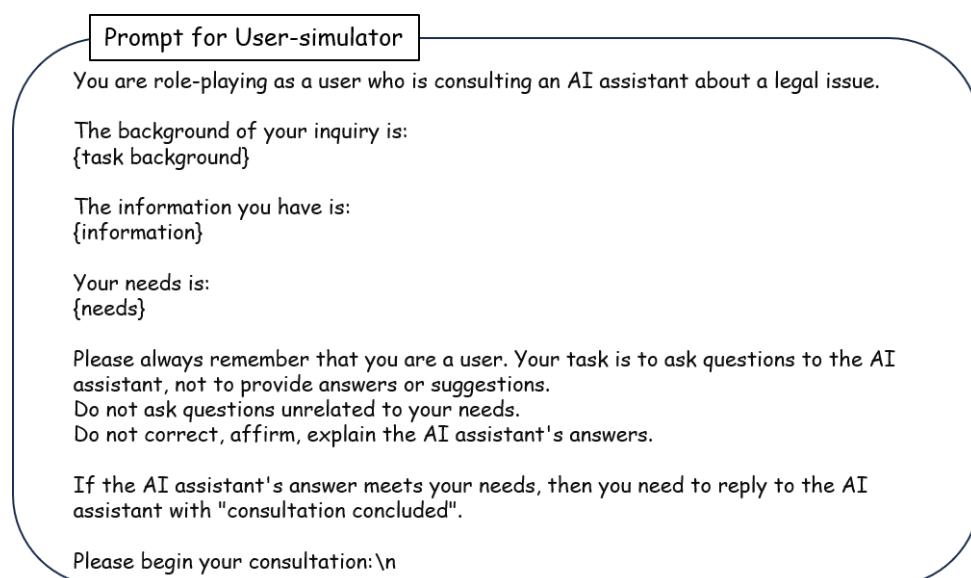


Figure 9: Translated version of user-simulator prompt template.

- **Qwen-Chat** Qwen series is trained on very large scale pre-training data. Pre-training data types are diverse and cover a wide range, including a large number of web texts, professional books, code, etc.
- **Qwen2.5** Qwen 2.5 excels in coding, mathematics, and long-text generation, supports structured data understanding and JSON output, and offers robust multilingual capabilities across 29 languages.
- **Internlm-Chat** Internlm is pre-trained on more than 2.3 terabytes of tokens containing high-quality English, Chinese, and code data. In addition, the Chat version has been trained

on SFT and RLHF, making it better and safer to meet users' needs.

- **Internlm2.5** InternLM2.5 has open-sourced a 7 billion parameter base model and a chat model tailored for practical scenarios.
- **Yi-Chat** The Yi series models are the next generation of open-source large language models trained from scratch by 01.AI, targeted as a bilingual LLM and trained on 3T multilingual corpus.

Legal LLMs

- **Wisdom-Interrogatory** Jointly developed by Zhejiang University, Alibaba Damo Insti-

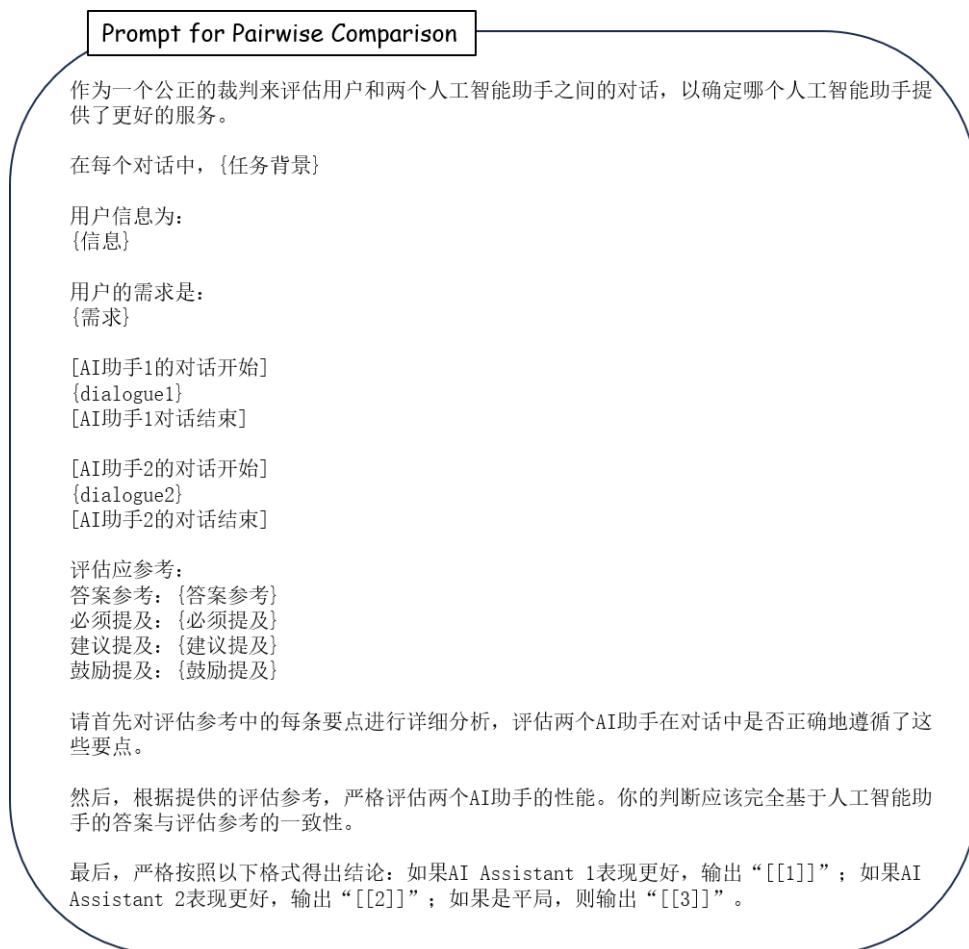


Figure 10: GPT-4 evaluation prompt template. The two dialogues were randomly shuffled and the LLMs’ name were hidden to eliminate position and name bias.

Model	Parameters	SFT	RLHF	Access	Base Model
ChatGPT	-	✓	✓	API	-
GPT-4	-	✓	✓	API	-
o1-mini	-	✓	✓	API	-
o1-preview	-	✓	✓	API	-
Baichuan2-7B-Chat	7B	✓	✓	Weights	Baichuan2-7B
Baichuan2-13B-Chat	13B	✓	✓	Weights	Baichuan2-13B
chatglm3-6b	6B	✓	✓	Weights	ChatGLM3-6B-Base
glm-4-9b-chat	9B	✓	✗	Weights	glm-4-9b
deepseek-llm-67b-chat	67B	✓	✗	Weights	deepseek-llm-67b-base
DeepSeek-V2-Chat	21B	✓	✓	Weights	DeepSeek-V2.5
Qwen-7B-Chat	7B	✓	✗	Weights	Qwen-7B
Qwen-14B-Chat	14B	✓	✗	Weights	Qwen-14B
Qwen-72B-Chat	72B	✓	✗	Weights	Qwen-72B
Qwen-2.5-7B-Instruct	7B	✓	✗	Weights	Qwen-2.5-7B
internlm-chat-7b	7B	✓	✗	Weights	internlm-7b
internlm-chat-20b	20B	✓	✓	Weights	internlm-20b
internlm-2.5-7b-chat	7B	✓	✓	Weights	internlm-2.5-7b
Yi-34B-Chat	34B	✓	✗	Weights	Yi-34B
Wisdom-Interrogatory	7B	✓	✗	Weights	Baichuan-7B
DISC-LawLLM	13B	✓	✗	Weights	Baichuan-13B-Base
fuzi-mingcha	6B	✓	✗	Weights	ChatGLM-6B

Table 10: List of LLMs

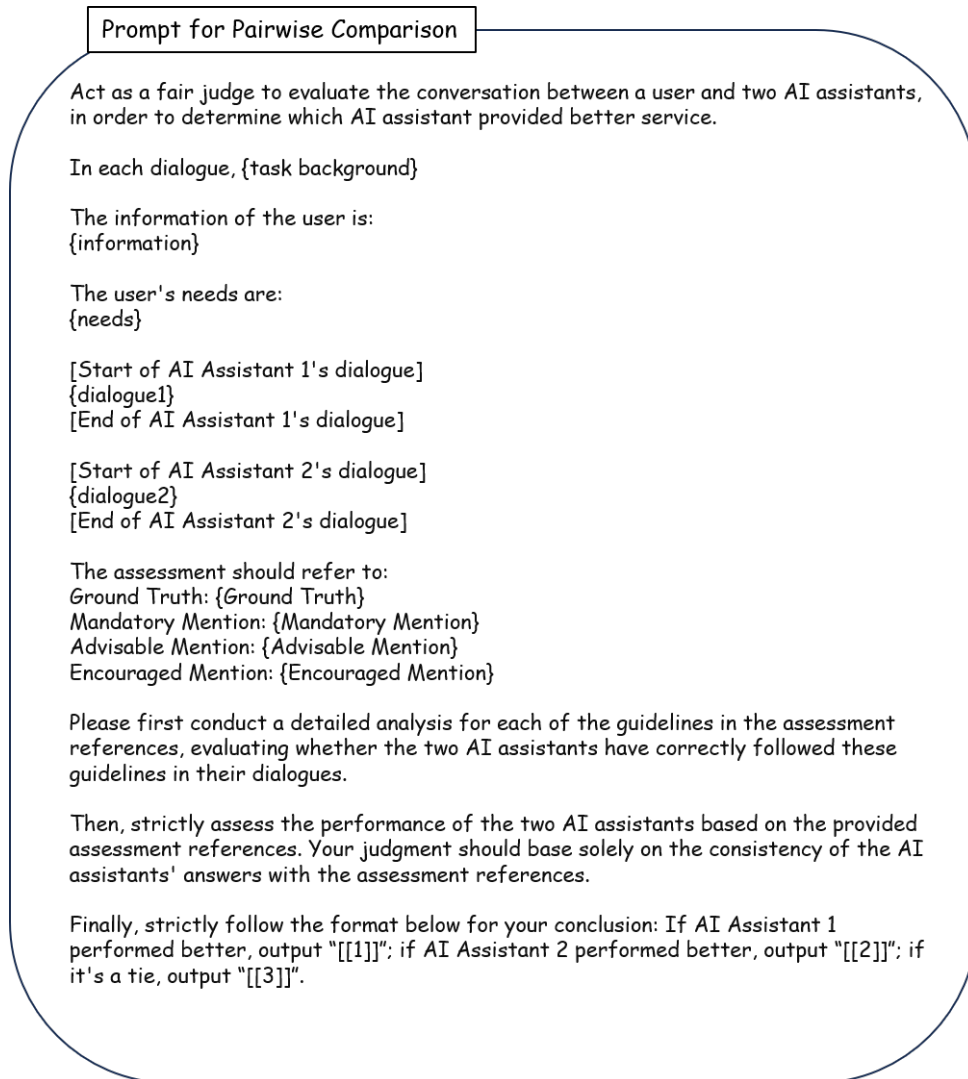


Figure 11: Translated version of GPT-4 evaluation prompt template.

Type of Answer Guidance	N	Points Referred Rate	Correct Assess Rate
Ground Truth	372	99.2%	85.9%
Mandatory Mention	510	98.0%	86.6%
Advisable Mention	1404	94.7%	74.4%
Encouraged Mention	1173	95.4%	84.9%
Summary	3459	95.9%	81.1%

Table 11: The rate of key points referred to and correctly assessed by GPT-4 during evaluation. 'Referred' means GPT-4 mentioned points from the answer guidance during the evaluation, while 'Correctly Assessed' means a human expert deemed that GPT-4 correctly judged whether two dialogues complied with each point.

tute and China Institute Computing. The base model is Baichuan-7B, which performs secondary pre-training and instruction fine-tuning. The pre-training data encompasses legal documents, judicial cases, and legal Q&A, amounting to 40 gigabytes. The fine-tuning dataset categories include legal examination questions, judicial consultations, legal scenario Q&A, prediction of legal violations and charges, sentencing prediction, court opinions,

and case summary extraction, with a total of 100k instances.

- **DISC-LawLLM** Developed by Fudan University, obtained by fine-tuning based on Baichuan-13B-Base. The supervised fine-tuning dataset DISC-Law-SFT includes legal information extraction, judgment prediction, document summarization, and legal question answering, with a total of 295k instances, en-

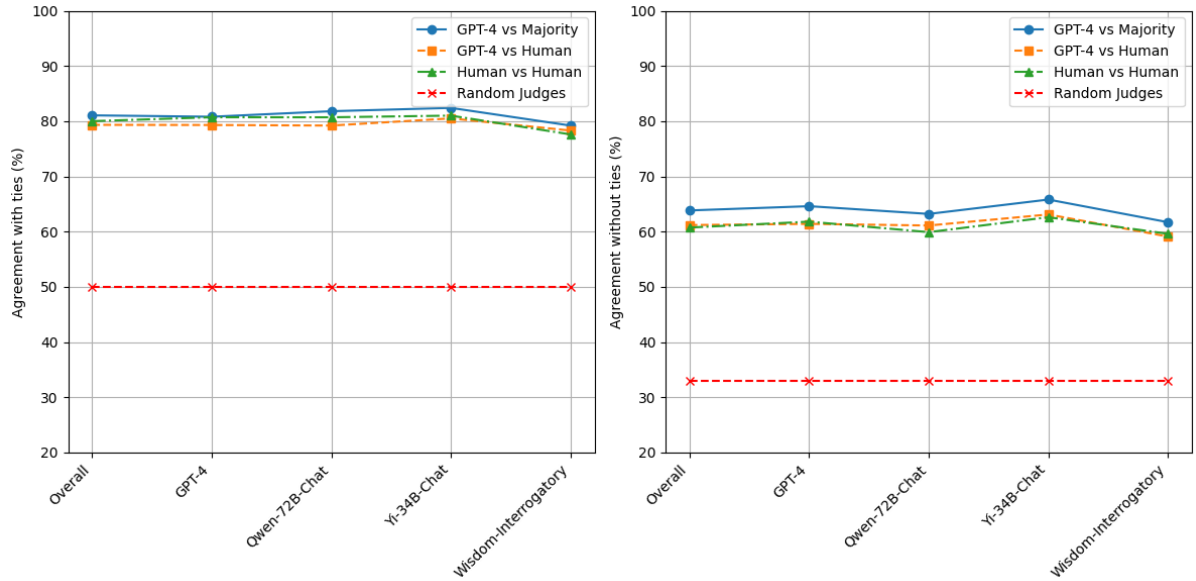


Figure 12: The agreement between GPT-4 and Human, GPT-4 and Majority, and Human and Human was assessed. The majority judgement was determined by majority voting among three annotators. The left sub-figure shows the agreement including both tie and no-tie samples, where annotators could mark a tie. The right sub-figure shows the agreement excluding ties, requiring annotators to make a clear preference among no-tie samples. Note that the agreements between two random judges are 50% and 33.3%, respectively.

sureing coverage of various judicial application scenarios. By fine-tuning this dataset, it focuses on providing legal service for various users like legal professionals general population and legal students.

- **fuzi-mingcha** Jointly developed by Shandong University, Inspur Cloud, and China University of Political Science and Law. It is based on ChatGLM and trained using a vast amount of unsupervised Chinese judicial corpus (including various judgment documents, laws, and regulations, totaling 27M) and supervised fine-tuning data for the judiciary (including legal Q&A, legal reasoning, similar case retrieval, totaling 154k instances). This model supports functions such as legal provision retrieval, case analysis, syllogistic reasoning for judgments, and judicial dialogue, aiming to provide users with comprehensive and highly accurate legal consultation and answer services.

General Performance of LLMs To provide a clearer context for the performance of the tested LLMs on UCL-Bench, we have added a summary Table 12 that outlines the general performance of these models on CMMLU, a general benchmark that assesses the capabilities of these models from

the perspective of Chinese knowledge and reasoning. We aim to clarify the relationship between general and legal capabilities, helping readers better understand the strengths and limitations of the evaluated models in both domains. The Pearson correlation coefficient between the models’ performance on general benchmarks and UCL-Bench is **0.50**, indicating a moderate positive correlation between general capabilities and legal-specific performance.

E.2 Analysis of GPT-4 judge’s Effectiveness

We first evaluate the effectiveness of GPT-4 judge from both process dimension and result dimension. In terms of process dimension, we assess whether GPT-4 judge referred to the answer guidance and made the correct judgments in its evaluation. In terms of result dimension, we calculate the agreement between human expert and the GPT-4 judge.

We also incorporate DeepSeek V2.5 as an evaluator for comparative analysis. This aims to build an adaptable evaluation framework, similar to AlpacaEval and WildBench, allows for the substitution of ChatGPT/GPT-4 with comparable models, ensuring the benchmark’s generalizability.

Process Dimension We extract 880 evaluation result samples of four LLMs: GPT-4, Qwen-72B-Chat, Yi-34B-Chat, and Wisdom-Interrogatory,

	DeepSeek-V2	Qwen-2.5 7B	GLM-4-9B	InternLM-2.5 7B	Baichuan2 13B	Baichuan2 7B	ChatGLM3 6B	Qwen 72B	Yi 34B
CMMLU	82.4	80.9	74.5	78.0	62.0	57.1	67.5	83.6	82.6
UCL-Bench	69.4	72.9	67.6	66.8	66.0	64.0	46.0	70.0	69.0

Table 12: Comparison of General and Legal-Specific Performance of LLMs

with 220 samples for each model. Three legal masters are involved to annotate the evaluation results given by GPT-4 judge with each paid 50 RMB per hour. For each annotated sample, they are provided with the prompt input of GPT-4 judge and the evaluation result it provides. They need to first calculate the proportion of key points from the Answer Guidance that GPT-4 referred to, and then determine whether GPT-4 made the correct judgment for each key point. To mitigate bias in the annotation process, we employed a majority voting mechanism. The results are shown in Table 11. On the one hand, the GPT-4 judge performs effectively in referring to the answer guidance, mentioning 95.9% of key points in the evaluation process. Specifically, for each category of answer guidance, when it contains fewer key points (such as Mandatory Mention), GPT-4 judge typically does not miss any of them. However, when a category includes multiple key points (such as Advisable Mention), it may sometimes overlook some points during the analysis. On the other hand, human experts deem that GPT-4 judge correctly determines whether the LLM’s answers align with the key points 80.4% of the time. Specifically, for points that are more objective like Ground Truth, Mandatory Mention, and Encouraged Mention, the agreement between human judges and GPT-4 is high, exceeding 84%. Conversely, for more subjective and longer points like Advisable Mention, the agreement is lower, at only 74.4%. Feedback from human experts also indicates that the boundaries for judging Advisable Mention content are blurred, making it challenging to provide accurate assessments.

Error Analysis To identify the types of mistakes made by GPT-4, we asked annotators to conduct further analysis on the error cases. The following primary error types were identified:

- **Missed Key Points in Evaluation**

In cases with numerous key points, GPT-4 occasionally missed some points during evaluation.

Case 1: In the Judgement Document Generation task, a specific data sample required the inclusion of four key points:

- Kidnapping with the intent to extort money;
- Recidivism;
- Prior criminal record;
- Adoption of the defense’s arguments.

However, during the evaluation, GPT-4 only assessed whether the first point was met. (GPT-4: “Both AI assistants clearly mentioned ‘kidnapping with the intent to extort money,’ fulfilling the requirement.”)

- **Preference for Detailed Responses Over Accuracy**

Despite explicit instructions in the prompt to disregard verbosity and comprehensiveness, GPT-4 occasionally favors detailed responses over accurate ones.

Case 2: In a specific data sample for the Judicial Exam Score task, the ground truth provided a scoring reference range of 4–5 points. AI Assistant 1 gave 4 points; AI Assistant 2 did not provide a score but was deemed to meet the key point due to its detailed analysis, even without a conclusion. (Evaluator: “Both AI assistants scored within the reasonable range, with AI Assistant 1 giving the user 4 points. Although AI Assistant 2 did not provide a specific score, it offered a more detailed analysis and identified specific scoring points. Thus, both AI assistants met the requirement in this regard.”)

- **Inaccurate Evaluation of Partially Met Key Points**

GPT-4 can only determine whether a key point is fully met, and its evaluation of partially met key points is often inaccurate.

Case 3: In a specific data sample for the Accusation Prediction task, the relevant legal provision should have been mentioned as Article 133, Section 1, Paragraph 1(b) of the Criminal Law. One AI assistant mentioned Article 133, Section 1, which should be considered partially correct. However, this was overlooked by the GPT-4 Evaluator. (Evaluator: “Neither AI assistant mentioned Article 133, Section 1,

Paragraph 1(b) of the Criminal Law, so their performance under this criterion was unsatisfactory.”)

Result Analysis The samples for annotation and the expert annotators are the same as mentioned above. However, during labeling, the evaluation results from the GPT-4 judge are not provided. The annotators’ task is to determine which LLM performed better based on the answer guidance. Similar to [Zheng et al.](#)’s approach, we conducted two sets of experiments: with ties and without ties. After the annotation is completed, we calculate the agreement between humans and GPT-4, as well as the agreement among humans. The results are shown in Figure 12. GPT-4 exhibits preferences in the evaluation process that closely resemble human majority, achieving an overall agreement rate of 81.1% in "with-out tie" setting and 63.4% in "with tie" setting. Furthermore, the agreement between GPT-4 and human annotators is very close to the inter-annotator agreement(80.1%, 60.7% respectively), highlighting GPT-4’s ability to replicate human-like judgment in dialogue assessment.

Overall, these results demonstrate that GPT-4 effectively utilizes answer guidance and aligns closely with human judgments, reinforcing its suitability as an effective judge in dialogue assessments.

Alternative Evaluator Comparing these results of Table 13 with those in original Table 3, we observe a consistent evaluation across both evaluators. This demonstrates that, within our framework, the evaluator is interchangeable and does not solely rely on GPT-4.

F Examples of Dialogue

We showcase multi-turn dialogues constructed from data samples to aid readers in better understanding our design philosophy. A total of 22 dialogues are presented, with each corresponds to a task, distributed across Tables 14 to 35. These dialogues are constructed through interactions between a user simulator and GPT-4.

	Article Recite	Legal Issue Consult	Judicial Interpret	Legal Concept Und.	Public Opinion Sum.	Judicial Sum.	Prison Term Pre.	Charge Predict	Accuse Predict	Article Predict	Dispute Focus Id.	Real Case Consult	Similar Case Retrieve	Legal Case Analysis	Judicial Exam Score	Judgment Doc. Pt.	Legal Essay Write	Cross-Exam. Gen.	Litigation Strategy Gen.	Judgment Doc. Gen.	Prosecutor Doc. Gen.	Lawyer Doc. Gen.	AVG
GPT-4	66.67	55.17	54.75	53.85	53.54	55.00	53.52	67.50	77.50	73.75	60.83	52.97	53.10	54.38	55.28	55.94	57.50	63.00	53.55	57.50	53.89	52.85	58.27
Qwen-72B	66.67	61.21	63.50	62.31	60.22	63.41	59.79	67.50	75.00	76.25	62.92	58.79	58.91	62.92	63.06	64.06	61.48	65.00	59.84	65.71	60.66	58.66	63.54
DISC-Law13B	56.67	43.10	48.25	44.62	42.78	45.91	42.97	53.75	55.00	61.25	50.83	40.14	41.34	46.67	46.94	49.69	50.71	52.00	43.23	50.71	43.80	40.47	47.77
Fuzi-Mingcha 7B	60.83	44.46	49.75	46.92	42.76	49.09	43.10	63.13	70.00	70.00	55.83	39.40	40.93	48.13	50.28	54.06	45.72	59.00	42.88	56.07	43.64	39.70	50.71

Table 13: Evaluation results of UCL-bench with the evaluator DeepSeek V2.5.

Information: None
Needs: 社会保险法第87条的内容是什么？(What is the content of Article 87 of the Social Insurance Law?)
User: 我不记得社会保险法第87条了。你能帮我快速搜索一下吗？(I can't remember Article 87 of the Social Insurance Act. Can you search it for me quickly?)
AI assistant: 是的，这是关于社会保险费的缴纳和使用的。具体文本如下： 第八十七条用人单位和个人应当依法参加社会保险，按照规定缴纳社会保险费。社会保险费由用人单位和个人按照规定的比例缴纳。社会保险费一经缴纳，按照国家规定用于支付社会保险待遇，不得挪作他用。“ 请注意，这是中国的社会保险法，与其他国家或地区的法律可能有所不同。 (Yes, this one is about the payment and use of social insurance premiums. The specific text is as follows: "Article 87: Employing units and individuals shall participate in social insurance according to law and pay social insurance premiums according to regulations. Social insurance premiums shall be paid by employers and individuals in accordance with the prescribed proportion. Social insurance premiums, once paid, shall be used for payment of social insurance benefits in accordance with state regulations and may not be diverted for other purposes." Please note that this is the social insurance law of China and the laws of other countries or regions may be different.)

Table 14: Example dialogue for Article Recitation(1-1).

Information: None
Needs: 在中国，70年后拆房子有赔偿吗？(Is there compensation for tearing down a house after 70 years in China?)
User: 在中国，房子的楼龄达到70年，如果被拆除，业主能得到补偿吗？(In China, when a house is 70 years old, if it is demolished, can the owner be compensated?)
<p>AI assistant: 在中国，房屋的所有权制度通常是70年，但这并不意味着70年后房产将归国家所有。实际上，这完全取决于当地政府的规定和程序。</p> <p>如果房屋被拆除，无论其年龄大小，权利持有人都有权获得相应的补偿，这是由政府规定的。包括但不限于土地补偿、房屋置换补贴或新房补偿。具体补偿的数额和方式需要根据具体城市、地区和土地的情况，以及政府的相关政策来确定。</p> <p>此外，建议在遇到此类法律问题时，最好咨询专业法人或直接与当地政府部门联系，以获得更准确和具体的信息。</p> <p>(In China, the ownership system of houses is usually 70 years, but this does not mean that the property will be owned by the state after 70 years. In practice, it all depends on local government regulations and procedures.</p> <p>If a house is demolished, regardless of its age, the right holder is entitled to corresponding compensation, which is regulated by the government. Including but not limited to land compensation, housing replacement subsidies or new housing compensation. The amount and method of specific compensation need to be determined according to the situation of the specific city, region and land, as well as relevant government policies.</p> <p>In addition, it is recommended that when encountering such legal issues, it is best to consult a professional legal person or contact the local government department directly for more accurate and specific information.)</p> <p>User: 在房屋拆迁的情况下，具体赔偿金额如何计算？(In the case of house demolition, how to calculate the specific compensation amount?)</p> <p>AI assistant: 在中国，房屋拆迁补偿具体金额的计算通常考虑以下因素：</p> <ol style="list-style-type: none"> 1. 土地使用权补偿费：补偿费的数额，通常参照该地区的地价，计算土地面积。 2. 房屋物质损失赔偿：根据房屋面积、结构、建筑成本等因素，按市场价格计算赔偿金额。 3. 搬迁费：包括搬迁费、临时住宿费等。 4. 维修费用：如果房子之前已经修理过，这部分费用将被考虑在内。 5. 房屋拆迁对业主正常生活和生产经营造成的影响：常见的如停工损失、营业损失等，通常需要提供相关证明材料证明损失。 <p>以上具体补偿金额需参照当地拆迁补偿政策和标准。</p> <p>(In China, the calculation of the specific amount of compensation for house demolition usually takes into account the following factors:</p> <ol style="list-style-type: none"> 1. Compensation for land use rights: The amount of compensation is usually calculated by referring to the land price of the area and calculating the land area. 2. Compensation for the material loss of the house: The compensation amount is calculated according to the market price based on the area, structure, construction cost and other factors of the house. 3. Relocation costs: including moving costs, temporary accommodation costs, etc. 4. Maintenance costs: If the house has been previously repaired, this part of the cost will be considered. 5. The impact of the owner's normal life and production and operation caused by house demolition: common such as stoppage loss, business loss, etc., usually need to provide relevant certification materials to prove the loss. <p>The above specific compensation amounts need to refer to the local demolition compensation policies and standards.)</p>

Table 15: Example dialogue for Legal Issue Consultation(2-1).

Information: 第一百四十一条: 生产、销售假药的, 处三年以下有期徒刑或者拘役, 并处罚金; 对人体健康造成严重危害或者有其他特别严重情节的, 处三年以上十年以下有期徒刑, 并处罚金; 造成他人死亡或者有其他特别严重情节的, 处十年以上有期徒刑、无期徒刑或者死刑, 并处罚金或者没收财产。吸毒者明知是假药而将其提供给他人使用的, 依照前款的规定处罚。

(Article 141: Whoever produces or sells counterfeit drugs shall be sentenced to fixed-term imprisonment of not more than three years or criminal detention and shall also be fined; If he causes serious harm to human health or if there are other particularly serious circumstances, he shall be sentenced to fixed-term imprisonment of not less than three years but not more than 10 years and shall also be fined; If he causes death to another person or if there are other especially serious circumstances, he shall be sentenced to fixed-term imprisonment of not less than 10 years, life imprisonment or death and shall also be fined or be sentenced to confiscation of property. Any person of a drug user who knowingly supplies a fake drug to another person for use shall be punished in accordance with the provisions of the preceding paragraph.)

Needs: 解释什么情况应被视为“其他特别严重的情节”。(Explain what circumstances should be considered "other particularly serious circumstances".)

User: 根据《中华人民共和国刑法》第一百四十一条的规定, 在什么情况下应当认定为“其他特别严重情节”? (According to Article 141 of the Criminal Law of the People's Republic of China, under what circumstances should it be identified as "other particularly serious circumstances"?)

AI assistant: “其他特别严重情节”的判断将根据具体案件的实际情况, 但通常包括以下情况: 大规模生产、销售假药; 生产和销售的假药质量极差, 对人体造成不可逆的严重危害; 受影响的人数特别多, 社会影响特别恶劣。这些情况可视为“其他特别严重情况”。但请注意, 这是司法解释, 具体适用要看具体情况和法官的裁决。

(The judgment of "other particularly serious circumstances" will be based on the actual circumstances of the specific case, but usually includes the following circumstances: large-scale production and sale of counterfeit drugs; The quality of counterfeit drugs produced and sold is extremely poor, causing irreversible and serious harm to the human body; The number of people affected is particularly large, and the social impact is particularly bad. These circumstances may be considered "other particularly serious circumstances". However, please note that this is a judicial interpretation, and the specific application depends on the specific circumstances and the ruling of the judge.)

Table 16: Example dialogue for Judicial Interpretation(2-2).

Information: 关于盗窃, 下列哪个说法是正确的?

A. 光华公司以B的名义登记A房, B谎称房产证遗失, 在不知情的情况下卖给C。B犯了入室行窃罪 B. 快递员A受快递公司委托收取“货到付款”货款, 隐匿部分货款, A构成盗窃

C. B来到A家, 喝醉了, A和B撬开了锁, 一起偷走了自行车。自行车被盗后, 后来才发现自行车是A, A构成盗窃

D. A在无人超市里, 把低价商品的标签贴在高价商品的标签上, 只付了低价, 就把高价商品拿走了。A构成盗窃

(Which of the following statements is true about theft?

A. Guanghua company registered A house in the name of B, B falsely claimed that the property certificate was lost, and sold it to C without knowing. B is guilty of burglary of a house

B. Courier A is entrusted by the Courier company to collect "cash on delivery" payment, concealing part of the payment, A constitutes theft

C. B came to A's home, got drunk, A and B picked the lock and stole the bike together. After the bicycle is stolen, it is later found that the bicycle is A, and A constitutes theft

D. A, in the unmanned supermarket, pasted the label of low-priced goods on the label of high-priced goods, only paid a low price, and took away the high-priced goods. A constitutes theft)

Needs: 回答这个问题。(Give the answer to the question.)

User: 以上问题的正确答案是什么? (What is the right answer to the above question?)

AI assistant: 这些问题的正确答案是B和D。

A项, B的行为构成欺诈, 不构成盗窃, 因为他以说谎的方式骗取他人财产为己用。

B项, A的行为属于侵占职务。他趁受快递公司委托代收货款的机会将钱偷走, 属于盗窃行为。

C项, A不占有他人的财产, 即使发生了盗窃, 也不构成盗窃, 因为盗窃的对象是A自己的财产。

D项, A的行为属于“以秘密窃取他人财物为目的的非法占有”, 构成盗窃罪。

(The correct answer to these questions is B and D.

Item A, B's conduct constitutes fraud rather than theft because he defrauded others of their property for his own use by lying.

Item B, the act of A belongs to the occupation of the position. He took advantage of the opportunity to steal the money when he was entrusted by the Courier company to collect the payment, which is theft.

Item C, A does not possess another person's property, even if the theft occurred, it does not constitute theft, because the object of theft is A's own property.

Item D, the conduct of A belongs to the "illegal possession for the purpose of secretly stealing the property of others", constituting the crime of theft.)

Table 17: Example dialogue for Legal Concept Understanding(2-3).

Information: 原标题: 拿刀威胁法官! 聊城一女子扰乱司法秩序被拘留15名梅某已被司法拘留
梅某携刀招报消息聊城12月23日讯 (记者袁志伟通讯员赵云) 12月23日, 聊城法院法警支队处理持刀威胁辱骂法官扰乱司法秩序事件。上午10时许, 离婚女子梅某在法院门口手持刀具哭闹并辱骂法官, 接到民警举报后, 法警支队紧急出动应急小组, 采取强制措施控制当事人, 收缴刀具, 并进行训诫, 梅某情绪激动不听警告, 扬言要结束与法官的案件处理。梅某的行为严重影响了正常的司法秩序, 造成了不良的社会影响, 被判处15日司法拘留。据悉, 梅某不服离婚退彩礼纠纷案判决, 该案法官反复宣读判决后释疑, 其仍不服法院判决, 于11月11日上午到法院门口长时间辱骂该法官, 该法官再次释疑后, 详细说明判决的法律适用; 并告知可以按照审判监督程序申请再审。梅某不听办案法官的解释、指导和劝阻; 11日中午和下午, 他又到法院门口长时间漫骂办案法官; 23日上午, 他持刀侮辱主审法官, 并采取拘留措施。聊城中院司法警察支队高度重视警务保护和日常安全保障工作, 加强司法警察正规化建设, 加强警力管理, 不断提高警察支队应对突发事件的能力, 为审判执行提供了有力的警务保障。

(Original title: Threaten the judge with a knife! Liaocheng a woman to disrupt the judicial order was detained 15 Mei Mou has been judicial detention Mei Mou carrying knives poster news Liaocheng December 23 news (reporter Yuan Zhiwei correspondent Zhao Yun) December 23, Liaocheng court bailiff detachment to deal with a knife threatening to abuse the judge to disrupt the judicial order incident. At 10 a.m., the divorce woman Mei Mou at the court gate carrying knives crying and verbally abusing the judge, after receiving the police information, the bailiff detachment emergency response team was dispatched urgently, take compulsory measures to control the parties, collect the knives, and admonish, Mei Mou emotional not to listen to the warning, threatened to end with the judge handling the case. Mei's behavior seriously affected the normal judicial order, resulting in adverse social impact, was sentenced to 15 days of judicial detention. It is reported that Mei Mou refused to accept the judgment of the divorce return bride price dispute case, the case judge repeated the judgment after the release of doubt, it still refused to accept the court decision, in the morning of November 11 to the court gate for a long time to abuse the case judge, the case judge again after the release of doubt, the legal application of the judgment in detail, and informed that it can apply for a retrial in accordance with the trial supervision procedure. Mei refused to listen to the explanation, guidance and dissuasion of the judge handling the case; At noon and in the afternoon of the 11th, he went to the gate of the court again for a long time to abuse the judge handling the case; On the morning of the 23rd, he insulted the presiding judge with a knife and took detention measures. Liaocheng Middle Court Judicial police detachment attaches great importance to police protection and daily security work, strengthens the regularization construction of judicial police, strengthens the management of police force, and continuously improves the ability of the police detachment to deal with emergencies, providing a strong police guarantee for the execution of trials.)

Needs: 为上述文章生成约100字的摘要。(Generate an abstract of about 100 words for the above article.)

User: 我需要100字左右的摘要, 概括文章中提到的梅某携带刀具扰乱司法秩序的事件, 包括梅某的行为、后果以及司法警察支队的应对措施。(I need a summary of about 100 words, which should summarize the incident mentioned in the article that Mei carried knives to disturb the judicial order, including her behavior, the consequences and the response measures of the judicial police detachment.)

AI assistant: 梅女士因拒绝接受离婚判决, 在法院门口持刀侮辱法官, 严重扰乱司法秩序。在司法警察支队紧急出动的情况下, 他们控制住了小梅, 收缴了小梅的刀具。梅的行为造成了恶劣的社会影响, 被依法判处15日司法拘留。梅对法官的解释和劝阻置若罔闻, 司法警察支队对此高度重视, 不断提高警务保护和日常安全保障能力。

(A divorcee surnamed Mei, who refused to accept the divorce judgment, insulted the judge at the gate of the court with a knife, seriously disturbing the judicial order. Under the emergency dispatch of the judicial police detachment, they controlled Mei and collected their knives. Mei's behavior caused a bad social impact and was sentenced to 15 days of judicial detention according to law. Mei turned a deaf ear to the judge's explanation and dissuasion, and the judicial police detachment attached great importance to it and constantly improved their police protection and daily security capabilities.)

Table 18: Example dialogue for Public Opinion Summary(2-4).

Information: 原告：赵某，男，汉族，永寿县人。被告：房，男，汉族，陕西省永寿县人。原告赵某诉被告房劳动合同纠纷一案，法院受理，依法组成合议庭，举行公开听证。原告到庭参加诉讼，被告公告后未到庭，法院依法进行了缺席审判，案件已经结案。原告方称：2016年9月25日至10月20日，被告聘请原告从事其承包的长宁镇至李家嘴高速公路肩路工程。干了一个月才付了部分款项，仍欠7200元未付，多次向被告催告，被告以各种理由拖欠至今。因此，诉讼请求：1。判决被告依法向原告支付劳动赔偿金7200元；2. 被告支付费用400元；3. 本案诉讼费由被告承担。被告在缺席的情况下没有回答。原告在法庭上提供了以下证据：一份劳动结算清单，证明被告欠原告劳动费用15170元。对于原告提供的上述证据，由于证据来源合法，内容确认，法院应当予以采纳。经审理查明：2016年9月25日至10月20日，被告人在原告承包的长宁镇至李家嘴公路肩路工程中施工。劳动补偿总金额为17170元，当时支付了2000元，一个月后支付了5000元，2017年腊月二十七日支付了3000元，尚欠7200元，双方结算清单可查。经原告多次催告被告后，被告以各种理由一直拖欠。法院认为，依法成立的合同对当事人具有法律约束力，当事人应当按照约定履行义务。本案中，原被告之间的劳动合同依法成立并有效，被告应及时偿还欠原告的劳动报酬，并支持原告要求赔偿7200元劳务费。原告要求的费用不予支持，因为没有提供相关证据。据此，根据《中华人民共和国民事诉讼法》第一百四十四条、《中华人民共和国民事诉讼法》第一百零八条和《中华人民共和国合同法》第八条、第六十条的规定，判决如下：自本判决生效之日起十日内，被告向原告赵某支付劳务费7200元，案件受理费50元由被告承担。在本判决确定的期限内不履行支付义务的，依照《中华人民共和国民事诉讼法》第二百五十三条的规定，延期履行期间债务的利息应当加倍计算。对本判决不服的，可以在收到判决书之日起15日内，向本院递交上诉状，并按对方当事人编号提交副本，向陕西省咸阳市中级人民法院提出上诉。

Plaintiff: Plaintiff Zhao sued defendant Room for a labor contract dispute. The court accepted the case, formed a panel, and held a public hearing. Zhao attended; Room did not and was tried in absentia. Zhao claimed that from September 25 to October 20, 2016, Room hired him for a highway shoulder project in Changning Town, owing 7,200 yuan in unpaid labor fees despite repeated requests. Zhao sought: 1) 7,200 yuan in labor compensation; 2) 400 yuan in fees; 3) legal costs. Room did not respond. Zhao provided a labor settlement list as evidence, showing Room owed 15,170 yuan. The court accepted the evidence as legal and confirmed. The court found that Zhao worked on the project from September 25 to October 20, 2016, with total labor compensation of 17,170 yuan. Room paid 2,000 yuan initially, 5,000 yuan a month later, and 3,000 yuan on December 27, 2017, leaving 7,200 yuan unpaid. The court ruled that the labor contract was valid and Room must pay the owed 7,200 yuan. Zhao's request for 400 yuan in fees was denied due to lack of evidence. According to the Civil Procedure Law, General Principles of Civil Law, and Contract Law, the court ordered Room to pay Zhao 7,200 yuan within ten days of the judgment's effective date and bear the 50 yuan case acceptance fee. Failure to pay would result in doubled interest under Article 253 of the Civil Procedure Law. Dissatisfied parties could appeal to the Xianyang Intermediate People's Court within 15 days.

Needs: 摘录法律文件摘要，用更短、更连贯、更自然的语言表达其主要内容。

(Extract a summary of the legal document and express its main content in shorter, more coherent, and natural language.)

User: 你能帮我摘录一下这份法律文件的主要内容，并做一个简短连贯的总结吗？(Could you help me extract the main contents of this legal document and give a brief and coherent summary?)

AI assistant: 原告赵某因劳动合同纠纷起诉被告房，欠其劳动赔偿金7200元。原告提供的劳动和解清单证据内容真实，法院确认原告方与被告之间的劳动合同合法有效。法院裁定被告应按期偿还欠款，因此法院支持原告要求支付7200元劳务费的请求，但原告未提供有关费用的相关证据，故不支持该费用请求。判决生效后10日内，被告应当向原告支付人民币7200元、人民币50元。被告不按时履行支付义务的，在迟延履行期间，应当依法加倍支付债务利息。(Plaintiff Zhao Mou sued defendant room due to labor contract dispute owing labor compensation 7,200 yuan. The evidence content of the labor settlement list provided by the plaintiff is true, and the court confirmed that the labor contract between the original and the defendant is legal and valid. The court ruled that the defendant should repay the arrears on time, so the court supported the plaintiff's request for 7,200 yuan of labor fees, but the plaintiff did not provide relevant evidence about the cost, the cost request was not supported. The defendant shall pay 7,200 yuan and 50 yuan to the plaintiff within 10 days after the judgment takes effect. If the defendant fails to fulfill the payment obligation on time, the interest on the debt during the delay in performance shall be doubled in accordance with the law.)

Table 19: Example dialogue for Judicial Summary(2-5).

Information: 唐山市丰润区人民检察院指控, 2013年8月5日21时, 在唐山市丰润区高里村集坪停车场浴池内, 被告人马某与陈某打牌后相互打斗, 随后陈某、马某、马某三人一起打斗, 打斗中, 马某用铁棒打伤了马某的左臂。经司法鉴定, 马某的伤情较轻, 陈某的伤情较轻。公诉机关就上述事实向法院提出了相应的证据。公诉机关认为, 被告人马某故意伤害他人, 造成轻伤, 应当以故意伤害罪追究刑事责任, 建议依照《中华人民共和国刑法》第二百三十四条第一款的规定量刑。

(Tangshan Fengrun District People's Procuratorate charged that at 21:00 on August 5, 2013, in the Jiping parking lot bath in Gaoli Village, Fengrun District, Tangshan City, the defendant Ma Mou and Chen Mou fought each other after playing cards, and then Chen Mou, Ma Mou, Ma Mou three people fought together, in the fight, Ma Mou injured Ma Mou's left arm with an iron rod. According to the judicial medical appraisal, the injury of Ma Mou was minor, and the injury of Chen Mou was minor. The public prosecution office submitted corresponding evidence to the court on the above facts. The public prosecution organ believes that the defendant Ma Mou intentionally injured others and caused minor injuries, should be investigated for criminal responsibility for intentional injury, and proposed to be sentenced in accordance with the provisions of the first paragraph of Article 234 of the Criminal Law of the People's Republic of China.)

Needs: 确定被告的刑期。(Determine the prison term of the accused.)

User: 根据被告马某某的行为及其后果, 在《中华人民共和国刑法》中, 他应该如何被判刑?(According to the defendant Ma XX's behavior and its consequences, in the Criminal Law of the People's Republic of China, how should he be sentenced?)

AI assistant: 被告马某的行为构成故意伤害罪。根据《中华人民共和国刑法》第二百三十四条第一款的规定, 故意伤害他人身体的, 处三年以下有期徒刑、拘役或者管制。鉴于马某行为造成的伤害程度为轻伤, 且为轻伤, 依法应给予较轻的处罚。具体量刑需要综合考虑被告人的前科、诉讼态度、悔过表示等因素, 但一般不会超过3年。

(The defendant Ma's behavior constitutes the crime of intentional injury. According to the provisions of Article 234, paragraph 1, of the Criminal Law of the People's Republic of China, whoever intentionally injures another person's body shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance. In view of the degree of injury caused by Ma's behavior is light injury, and minor injury, according to law should be given a lighter punishment. The specific sentence needs to be assessed based on a combination of factors such as the defendant's prior criminal record, attitude toward litigation, and expression of remorse, but generally speaking, it will not exceed three years.)

Table 20: Example dialogue for Prison Term Prediction(3-1).

Information: 公诉机关指控, 2015年5月中旬至7月, 被告人李某利用其在上海市宝山区大华路XXX水果店的工作机器, 10余次从店主吴某二楼卧室床下的鞋盒中盗取受害人销售资金共计1.5万余元。被告人李某再次被盗窃时, 上述事实当场被追回, 被告人李某在庭审中也没有提出异议, 且有被害人吴某的供述、上海市公安局宝山分局出具的《工作情况》、被害人出具的《谅解书》、被告人李某的供词及指认记录等证据确认, 足以认定。

(The public prosecution organs charged that from mid-May to July 2015, the defendant Li Mou used his working machine in the XXX fruit shop, Dahua Road, Baoshan District, Shanghai, to steal the victim's sales funds totaling more than 15,000 yuan from the shoebox under the bed in the second-floor bedroom of the owner Wu Mou more than 10 times. The defendant Li Mou was stolen again when the above facts were recovered on the spot, the defendant Li mou also had no objection in the trial, and there was a statement by the victim Wu mou, the "work situation" issued by the Shanghai Public Security Bureau Baoshan Branch, the "Letter of understanding" issued by the victim, the defendant Li mou's confession and identification records and other evidence confirmed, enough to identify.)

Needs: 给出被告的指控。(Give the defendant the charge.)

User: 根据控方的说法, 被告应该被控什么罪名?(According to the prosecution, what should the accused be charged with?)

AI assistant: 根据控方的说法, 被告应被指控犯有盗窃罪。(According to the prosecution, the defendant should be charged with theft.)

Table 21: Example dialogue for Charge Prediction(3-2).

Information: 经审查，查明被告董富于2022年7月21日21时许在北京市怀柔区**镇**村餐厅因鸡毛小事与被害人孔XX发生口角，被告董富用椅子击打孔XX头部，致使孔XX头部受伤。经鉴定，被害人孔某身体损伤程度属于轻伤级别。被告人董富于2023年5月8日被警方逮捕，并如实供述上述事实。确认上述事实的证据如下：被告人供词与辩护：被告人董富的供词与辩护；2. 受害人陈述：受害人孔的陈述；3. 证人证言：证人毕、张的证言；4. 鉴定意见：北京市怀柔区公安司法鉴定中心；5. 检查、检查、鉴定记录：对被告人董富的记录进行现场检查、鉴定的记录；6. 视听资料：现场监控录像；7. 其他证明材料：报警收据、案件受理登记表、到货历史、户籍证明、网上比对工作记录等。上述证据收集程序合法，内容客观真实，足以确定指控事实。被告人董富对所指控的犯罪事实和证据无异议，主动认罪。(According to the examination in accordance with the law, it was found that the defendant Dong Fu had a quarrel with the victim Kong XX due to trivia in the restaurant of ** Village, ** Town, Huairou District, Beijing at about 21:00 on July 21, 2022, and the defendant Dong Fu beat Kong XX on the head with a chair, resulting in a head injury to Kong XX. After identification, the victim Kong so-and-so body injury degree belongs to the minor injury level. The defendant Dong Fu was arrested by the police on May 8, 2023, and then truthfully confessed the above facts. The evidence for ascertaining the above facts is as follows: 1. Defendant's confession and defense: Defendant Dong Fu's confession and defense; 2. Victim statement: Victim Kong's statement; 3. Witness testimony: the testimony of the witnesses Bi and Zhang; 4. Appraisal opinion: Beijing Huairou District Public Security Judicial appraisal Center; 5. Records of inspection, inspection and identification: records of on-site inspection and identification of defendant Dong Fu's records; 6. Audiovisual materials: on-site surveillance video; 7. Other certification materials: alarm receipt, case acceptance registration form, arrival history, household registration certificate, online comparison work record, etc. The above evidence collection procedure is lawful, the content is objective and true, enough to determine the facts of the accusation. The defendant Dong Fu has no objection to the alleged criminal facts and evidence, and voluntarily pleaded guilty.)

Needs: 给出对被告的指控和相关法律。(Give the charges against the accused and the relevant law.)

User: 请分析被告被指控的罪名，并提供相关法律内容。(Please analyze what the defendant is charged with and provide the relevant legal content.)

AI assistant: 根据被告的犯罪事实，其行为已构成故意伤害罪。《中华人民共和国刑法》第二百三十四条规定：“故意伤害他人身体的，处三年以下有期徒刑、拘役或者管制。”由于鉴定结果表明被害人的伤害为轻伤，可以按照本条的规定进行处理。

另外，对于被告人供述的处理，第六十七条规定：“犯罪后自动自首，如实供述的，为自首人。”自首的，可以从轻或者减轻处罚。主动赔偿被害人的损失，或者取得被害人的谅解，消除或者减轻损害后果的，可以减轻处罚。”

本案中，被告人董富在被逮捕后，如实供述了自己的犯罪行为，可以视为自首，对其量刑可能有一定的减刑作用，但最终的量刑需要法院根据各种情况判断。

(According to the defendant's criminal facts, his behavior has constituted the crime of intentional injury. Article 234 of the Criminal Law of the People's Republic of China stipulates: "Whoever intentionally injures another person's body shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance." Since the results of the identification show that the victim's injury is a minor injury, it can be treated in accordance with this article.

In addition, regarding the handling of the defendant's confession, Article 67 stipulates: "A person who automatically surrenders himself after committing a crime and confesses his crime truthfully is a person who surrenders himself." A person who has surrendered himself may be given a lighter or mitigated punishment. If a person actively compensates for the loss of the victim, or obtains the understanding of the victim, and eliminates or mitigates the consequences of the damage, the punishment may be mitigated."

In this case, after the defendant Dong Fu was arrested, he truthfully confessed his criminal behavior, which can be regarded as a surrender, which may have a mitigating effect on his sentence, but the final penalty needs to be judged by the court according to all the circumstances.)

Table 22: Example dialogue for Accusation Prediction(3-3).

Information: 2021年11月7日21时05分, 被告范驾驶一辆小型汽车, H×××**, 从本市翔一明园小区出发, 经水斗口大道、合肥路、深圳路, 沿深圳路自西向东至深圳路与宁联路交叉口西侧50米处, 与护栏相撞, 造成车辆及护栏损坏的道路交通事故。警方在巡逻时当场查获了这枚戒指。经鉴定, 被告人范某血液中乙醇含量为210.2mg/100ml。

(At 21:05 on November 7, 2021, Defendant Fan drove a small car, No. H×××**, from the Xiangyi Mingyuan Community in this city, through Shuidukou Avenue, Hefei Road and Shenzhen Road, along Shenzhen Road from west to east to the west side of the intersection of Shenzhen Road and Ninglian Road 50 meters, collided with the guardrail, causing a road traffic accident in which the vehicle and the guardrail were damaged. It was seized at the scene by police on patrol. After identification, the ethanol content of defendant Fan's blood was 210.2mg/100ml.)

Needs: 确定这个案件涉及的法律条文。(Determine the legal provisions of the case.)

User: 本案被告范某涉嫌触犯哪些法律规定?(What legal provisions is the defendant Fan suspected of in this case?)

AI assistant: 被告范的行为涉嫌下列法律规定:

(一) 《中华人民共和国刑法》第一百三十三条规定:“醉酒驾驶机动车, 或者驾驶机动车时饮酒, 或者发生重大交通事故逃逸, 尚不构成犯罪的, 吊销机动车驾驶证, 五年内不得重新取得机动车驾驶证。”醉酒驾驶机动车的, 依照前款的规定处罚, 处拘役。”

(二) 道路交通安全法第二十九条:“驾驶人违反道路交通安全法, 违反驾驶证, 有下列情形之一的, 处十五日以下拘留; 构成犯罪的, 依法追究刑事责任: (1) 醉酒驾驶机动车的; ”

(三) 《道路交通安全法》第二十六条规定:“驾驶机动车在道路上行驶, 应当遵守交通信号、标志、标线的规定, 保证安全、秩序。”

虽然这只是一起道路损毁和车辆碰撞护栏的事故, 但驾驶员不仅醉酒驾驶车辆, 而且还在道路上行驶, 构成了严重的交通违法行为。

Defendant Fan's conduct is suspected of the following legal provisions:

(1) Article 133 of the Criminal Law of the People's Republic of China reads: "If a person drives a motor vehicle after being drunk, or drinks alcohol while driving a motor vehicle, or escapes after a major traffic accident, and no crime has been constituted, his motor vehicle driving license shall be revoked and he shall not be allowed to re-obtain his motor vehicle driving license within five years." Whoever drives a motor vehicle after being drunk shall be punished in accordance with the provisions of the preceding paragraph and be sentenced to criminal detention."

(2) Article 29 of the Road Traffic Safety Law: "A driver (under any of the following circumstances) shall be detained for a maximum of 15 days for violating the road traffic Safety Law and driving permit; If a crime has been constituted, criminal responsibility shall be investigated according to law) : (1) driving a motor vehicle while intoxicated;"

(3) Article 26 of the Road Traffic Safety Law: "When driving a motor vehicle on a road, it shall comply with the provisions of traffic signals, signs and line marking to ensure safety and order."

Although it is only a road damage and vehicle collision guardrail incident, the driver is not only drunk driving the vehicle, but also driving on the road, which constitutes a serious traffic violation.

Table 23: Example dialogue for Article Prediction(4-1).

Information: ORG对原判决不服, 向法院提出上诉, 请求法院依法变更判决, 其事实和理由是: 上诉人在原判决中承担的伤残赔偿金数额违法。被上诉人PER的劳动合同签订日期为TIME, 事故发生日期为TIME, 其工作时间小于事故发生时的时间, 被上诉人PER家庭账户性质为有地农户账户。根据相关法律, 农村人口只有在城市连续居住或工作, 才能按城市标准支付。被上诉人PER不在此种情况下, 也没有资格根据城镇标准获得赔偿。被上诉人答辩: 原审认定事实清楚, 适用法律正确, 请求维持原判。

(The original judgment ORG refused to accept, to appeal to the court, request the court to amend the judgment according to law, the facts and reasons are: the amount of disability compensation borne by the appellant in the original judgment is illegal. The labor contract of the Appellee PER was signed on the date of TIME, the accident occurred on the date of TIME, and its working TIME was less than time when the accident occurred, and the nature of the Appellee PER household account was an agricultural household account with land. According to the relevant laws, the rural population can be paid according to the urban standard only if they live or work continuously in the city. Appellee PER is not in this situation and is not eligible for payment according to town standards. Appellee PER defense: The original trial found the facts clearly, the application of the law is correct, request to maintain the original judgment.)

Needs: 给出争议焦点。(Give the dispute focus.)

User: 根据本案的资料, 你能确定出争议焦点吗?(Based on the information of this case, can you determine the main point of dispute?)

AI assistant: 是的, 从案件资料来看, 争议的焦点主要有两个方面: 一是上诉人对原审裁定的伤残赔偿金数额的合法性提出质疑; 二是上诉人因工作时间不足和户籍性质是否符合城市补偿标准的争议。(Yes, according to the information of the case, the focus of the dispute is mainly two aspects: First, the appellant's challenge to the legality of the amount of disability compensation awarded in the original trial; The second is the dispute over whether the appellant meets the urban standards for compensation due to insufficient working hours and the nature of the household registration.)

Table 24: Example dialogue for Dispute Focus Identification(4-2).

Information: 合同上的原始印章不清楚。(The original seal on the contract is not clear.)
Needs: 合同可以重新盖章吗? (Can the contract be stamped again?)
User: 我有一份合同, 上面的盖章不是很清楚, 我可以重新盖章吗? (I have a contract, the stamp on it is not very clear, can I stamp it again?)
AI assistant: 如果已经盖章但不清楚, 可以重新盖章, 但最好在原盖章的地方重新盖章, 以免造成合同被故意修改或篡改的误解。最重要的是, 你需要让另一方同意再次盖章, 以避免任何可能的纠纷。 然而, 最好的建议是直接咨询相关的缔约方或法律顾问。在许多情况下, 您将需要第三方证人或证明您的重新盖章是由于缺乏清晰度, 而不是更改或篡改合同。 (If you have already stamped it but it is not clear, you can do it again, but it is best to re-stamp it in the place of the original seal to avoid the misunderstanding that the contract is deliberately modified or tampered with. Most importantly, you need to get the other party to agree to stamp again to avoid any possible disputes. However, the best advice is to consult directly with the relevant contracting party or legal counsel. In many cases, you will need to have a third party witness or certify that your reseal was due to a lack of clarity, rather than an alteration or tampering with the contract.)
User: 如果我没有得到对方的同意, 我可以重新盖章吗? (If I do not have the other party's consent, can I stamp again?)
AI assistant: 一般来说, 未经另一方同意更改合同, 包括重新盖章, 都可能被视为非法。这种行为可能会使对方认为你在更改合同内容或篡改信息, 从而引发法律纠纷。因此, 如果你打算重新盖章, 一定要得到另一方的明确许可, 最好有第三方在场见证。如果不能获得同意, 你应该尽快联系律师, 寻求专业的法律建议。 (In general, altering a contract without the other party's consent, including resealing it, may be considered illegal. This behavior may cause the other party to think that you are changing the contents of the contract or tampering with information, which can lead to legal disputes. Therefore, if you intend to re-stamp, be sure to obtain the express permission of the other party, and preferably a third party is present to witness. If consent cannot be obtained, you should contact an attorney for professional legal advice as soon as possible.)
User: 谁应该是第三方证人? (Who should be the third party witness?)
AI assistant: 选择作为第三方证人的人通常应该具有一定程度的公正性和可信度。在许多情况下, 律师、公证人、调解员或其他类似公正的专业人员可以作为第三方证人。在某些情况下, 你也可以选择朋友或家人作为证人, 但要注意, 他们可能无法保证绝对的公正性, 所以一般不建议在这种情况下。 无论是谁, 关键是谁这个人将能够在必要时验证您只是重新密封了原始印章, 因为它不清楚, 并且没有对合同进行任何更改。在法律上, 这样的证词可以被视为证据。 (A person who chooses to be a third party witness should usually possess a certain degree of impartiality and credibility. In many cases, a lawyer, notary public, mediator, or other similarly impartial professional may serve as a third party witness. In certain cases, you can also choose friends or family members as witnesses, but be aware that they may not be able to guarantee absolute impartiality, so it is generally not recommended in such cases. Whoever it is, the key is that this person will be able to verify if necessary that you simply resealed the original seal because it was not clear and that no changes have been made to the contract. In law, such testimony can be considered evidence.)

Table 25: Example dialogue for Real Case Consultation(4-3).

Information: 原始案件: 原始案件 (Original case: Original case) 类案1:{类案1 } (Relevant case I:{Relevant case I}) 类案2:{类案2} (Relevant case II:{Relevant case II})
Needs: 指出哪个类案与原始案件更相似。(Indicate which case in question is more similar to the original case.)
User: 哪个案件与原始案件更相似? (Which case is more similar to the original case?)
AI assistant: 案件1与原始情形更相似。在这两个案件中, 都是针对贷款的法律案件, 但在案件1中, 原告与被告之间的互动、贷款协议的内容以及还款的情况(被告未能按时偿还贷款, 原告多次未能催收)与原案非常相似。虽然贷款涉及的金额和利率不同, 但案件的核心问题和基本事实是相同的, 即借款人未能按时偿还贷款, 原告多次未能催收, 最终将案件告上法院。因此, 所讨论的案件1与原始案件更相似。 (Case 1 is more similar to the original case. In the two cases concerned, both were legal cases against the loan, but the interaction between the plaintiff and the defendant in case 1, the content of the loan agreement and the circumstances of repayment (the defendant failed to repay the loan on time, the plaintiff repeatedly failed to collect) were very similar to the original case. Although the amount and interest rate involved in the loan are different, the core problem and basic facts of the case are the same, that is, the borrower failed to repay the loan on time, the plaintiff repeatedly failed to collect, and finally brought the case to the court. Thus, case 1 in question is more similar to the original case.)

Table 26: Example dialogue for Similar Case Retrieval(4-4).

Information: 案例：林某知道洪某（已被判刑）犯罪，向洪某提供了一个网络游戏网站，洪某将玩家重新加载到游戏中的1.2万元转账。同时，林某知道张某（已被判处刑事处罚）有犯罪行为，并向其提供了钓鱼网站，张某盗取了该玩家的账号和密码，并将账户中的资金转移。经A市B区公安局调查，认为林某在明知他人利用信息网络犯罪的情况下，仍为他人提供网站，构成从犯，将案件移送B区检察院审查起诉，由B区检察院向B区法院提起公诉。本案证据包括：价格确认机构出具的价格确认书、林某的供词、讯问林某时的同步音视频录音、具有专业知识的人对鉴定意见的意见。

(Case: Lin knew that Hong (who has been sentenced to criminal punishment) committed a crime, and provided him with an online game website, and Hong transferred the 12,000 yuan that players reloaded into the game. At the same time, Lin knew that Zhang (who has been sentenced to criminal punishment) committed a crime, and provided him with a phishing website, and then Zhang stole the player's account and password, and transferred the funds in the account.

After investigation by the Public Security Bureau of District B of A City, it is believed that Lin Mou, knowing that others use information networks to commit crimes, still provides websites for them, constituting an accessory offender, and transferred the case to the Procuratorate of District B for review and prosecution, and then the procuratorate of District B filed a public prosecution to the court of District B.

The evidence in this case includes: the price confirmation letter issued by the price confirmation agency, Lin's confession, the synchronous audio and video recording of interrogating Lin at a certain time, and the opinions of people with specialized knowledge on the appraisal opinions.)

Needs: 本案属于利用计算机实施犯罪，如何确定利用计算机实施犯罪的管辖权？(This case belongs to the use of computers to commit crimes, how to determine the jurisdiction of using computers to commit crimes?)

User: 在中国，利用电脑犯罪的案件应该由哪个司法机关负责？

(In China, which judicial organ should be in charge of the case of using computers to commit crimes?)

AI assistant: 在中国，计算机犯罪案件先由公安机关侦查，然后移送检察院审查起诉，最后由法院审判。其中，公安、检察院、法院都是有管辖权的司法机关。在具体划分上，公安部门是刑侦机关，检察院是公诉机关，法院是审判机关。因此，本案中的行为在刑事诉讼过程中分别由公安局、检察院和法院处理。

(In China, cases of computer crimes are first investigated by the public security organs, then transferred to the procuratorate for examination and prosecution, and finally tried by the courts. Among them, public security departments, procuratorates, and courts are all judicial organs with jurisdiction. In the specific division, the public security department is the criminal investigation organ, the procuratorate is the public prosecution organ, and the court is the trial organ. Therefore, the actions in the case are dealt with by the public Security Bureau, the procuratorate and the court respectively in the course of criminal proceedings.)

Table 27: Example dialogue for Legal Case Analysis(4-5).

Information: (该题目满分为4分) 案件：赵某的妻子张某与被告马某的父亲马某因鸡毛蒜皮的小事发生争执，赵某将张某手中的汤饭扔向马某一脸。被告人马某得知后，将马某一及其兄弟马某二叫到赵某家门前辱骂，并与赵某及其兄弟郭某发生斗殴。在此过程中，马某误伤郭某，后经鉴定为轻伤，并致赵某死亡。

问题：张某是否有权提起附带民事诉讼？

我的回答：没错。本案中，受害人赵某已经死亡，其妻子张某作为赵某的近亲属，有权提起附带民事诉讼。该法的依据是《刑事诉讼法》第101条第1款。

(The full score of the subject is 4 points)Case: Zhao mou's wife Zhang Mou and the defendant Ma mou's father Ma Mou due to trivia dispute, Zhao Mou will Zhang hands soup rice thrown to Ma Mou a face. After the defendant Ma Mou learned this, he called Ma Mou I and his brother Ma Mou II to abuse him in front of Zhao mou's house, and fought with Zhao mou and his brother Guo mou. In the process, Ma Mou accidentally injured Guo mou, which was later identified as a minor injury, and killed Zhao Mou. Question: Does Zhang have the right to bring an incidental civil action?

My answer: Right to mention. In this case, Zhao, one of the victims, has died, and his wife Zhang, as a close relative of Zhao, has the right to file an incidental civil lawsuit. The law is based on Article 101, paragraph 1, of the Criminal Procedure Code.)

Needs: 这道题满分是4分，想知道我的答案能得几分？(This question is full of 4 points, want to know my answer can score?)

User: 请问在满分为4分的情况下，我的答案可以得多少分？(May I ask how many points can my answer get in the case of a full score of 4 points?)

AI assistant: 你的回答是正确的，对刑事附带民事诉讼权的理解和运用有充分的法律依据。所以，你得到了满分：4分。(Your answer is correct and has sufficient and legal basis for the understanding and use of incidental civil action rights in criminal cases. So, you get a perfect score: 4.)

Table 28: Example dialogue for Judicial Examination Scoring(5-1).

Information:	“根据合同法，如果租赁协议中没有明确规定，一般认为租金应支付由承租人给房东。” ("According to contract law, if the lease agreement does not explicitly specify, it is generally considered that the should rent be paid by the tenant to the landlord.")
Needs:	校对上述法律文件摘要。(Proofread the above legal document snippet.)
User:	请帮我校对一下以上法律文件的摘要。(Please help me proofread the above legal document snippet.)
AI assistant:	经校对，建议将文件的修改片段修改如下：“根据合同法，如果租赁协议没有明确规定，一般认为租金应由承租人支付给房东。”After proofreading, the revised snippet of the document is suggested as follows: "According to contract law, if the lease agreement does not explicitly specify, it is generally considered that the should rent be paid by the tenant to the landlord."

Table 29: Example dialogue for Judgment Document Proofreading(5-2).

Information:	问题：结合在法治轨道上统筹做好疫情防控工作的要求，探讨法治在国家治理体系和治理能力现代化中的积极作用。要求:1。没有观点或论点的，抄袭原创材料不计分；2. 观点正确，表述完整准确；3. 全文不少于500字。(Question: In combination with the requirements of coordinating epidemic prevention and control work on the track of rule of law, discuss the positive role of rule of law in the modernization of national governance system and governance capacity. Requirements: 1. There is no point of view or argument, copying the original material will not be scored; 2. Correct views, complete and accurate statements; 3. No less than 500 words in total.)
Needs:	给出司法考试中上述论述题的答案。(Give the answers to the above essay questions in the judicial examination.)
User:	给出司法考试中上述论述题的答案。(Give the answer to the above essay questions in the judicial examination.)
AI assistant:	<p>法治是治国理政的重要手段和框架。在疫情防控工作中，通过法律法规规范相关行为，为社会行为提供可操作的法律依据，营造有序明晰的社会环境。加大对危害疫情防控行为的执法司法力度，严格执行相关法律法规，是用法律之剑打好疫情防控阻击战，也是提高社会整体应对疫情能力的重要途径。</p> <p>事实上，这种以法治为基础的治理体系现代化，不仅在疫情防控中发挥着重要作用，而且在我们社会的各个方面都发挥着重要作用。法治保障人民基本权益，维护社会公平正义。它是治理体系的基础。坚实的法治基础，为我国治理体系和治理能力现代化提供了有力支撑。</p> <p>从广义上看，法治在推进我国国家治理体系和治理能力现代化的过程中，实际上促进了社会制度的完善和优化。依法治国方式在包括但不限于疫情防控在内的各个层面、各个环节发挥作用。这也有利于不断完善社会制度，为实现中华民族伟大复兴提供坚实保障。</p> <p>综上所述，法治在我国国家治理体系和治理能力现代化中的作用是全方位、多层次、深远的。我们必须珍惜法治的这种作用，最大限度地发挥法治在治理中的积极作用，为我国现代化进程的长治久安提供有力保障。</p> <p>(The rule of law is an important means and framework of governance. In epidemic prevention and control work, relevant behaviors are regulated through laws and regulations, providing operable legal basis for social behavior and creating an orderly and clear social environment. Increasing law enforcement and judicial efforts against behaviors that endanger epidemic prevention and control and strictly implementing relevant laws and regulations is a way to fight the epidemic with the legal sword and can enhance society's overall ability to respond to the epidemic.</p> <p>In fact, the modernization of this governance system based on the rule of law not only plays an important role in epidemic prevention and control, but also plays an important role in all aspects of our society. The rule of law protects the basic rights and interests of the people and maintains social fairness and justice. It is the foundation of the governance system. The solid foundation of the rule of law provides strong support for the modernization of our country's governance system and governance capabilities. From a broad perspective, the rule of law actually promotes the improvement and optimization of the social system in modernizing our country's national governance system and governance capabilities. The rule of law-oriented governance method plays a role at all levels and links, including but not limited to epidemic prevention and control. This role also helps to continuously optimize our social system and provides a solid guarantee for the great rejuvenation of the Chinese nation.</p> <p>To sum up, the role of the rule of law in the modernization of my country's national governance system and governance capabilities is all-round, multi-level, and far-reaching. We must cherish this role of the rule of law, maximize the positive role of the rule of law in governance, and provide a strong guarantee of lasting stability for our country's modernization process.)</p>

Table 30: Legal Essay Writing(6-1).

Information: 2013年6月20日，叶海群、赵树斌、叶海库、叶颖、马中山5人进入吉林大成广盛塑料工业有限公司施工现场从事电气施工布线工作，10月15日离开施工现场。吉林大成广盛塑料工业有限公司将该项目承包给自然人李军。该项目欠叶海群、赵树斌、叶海库、叶颖、马中山五人工资6.5万元。

被申请人叶海群向本院提交了以下证据：

一份叶俳句手写的水电报告复印件可以证明，五名被告都在申请人工施工。

(Five people, Ye Haiqun, Zhao Shubin, Ye Haiku, Ye Ying, and Ma Zhongshan, entered the construction site of Jilin Dacheng Guangsheng Plastic Industry Co., Ltd. on June 20, 2013, to engage in electrical construction wiring work, and left the construction site on October 15. Jilin Dacheng Guangsheng Plastic Industry Co., Ltd. contracted the project to Li Jun, a natural person. The project owed five people, Ye Haiqun, Zhao Shubin, Ye Haiku, Ye Ying and Ma Zhongshan, 65,000 yuan in wages.

Respondent Ye Haiqun submitted the following evidence to this court:

A copy of the water and electricity report handwritten by Ye Haiku proves that the five respondents are all applying for artificial construction.)

Needs: 对被申请人叶海群的证据提出质证意见。(Give the cross-examination opinions on the evidence of the respondent Ye Haiqun.)

User: 对于叶海群提供的叶海库手写的水电报告副本，我们该如何质证？(How should we cross-examine the copy of the water and electricity report handwritten by Ye Haiku provided by Ye Haiqun?)

AI assistant: 对于叶海群提供的叶海库水电手写报告副本质证，我建议提出以下问题进行质疑：

1. 关于叶俳句手写的水电报告，是否有任何工作记录、工作日志或合同形式的正式文件来支持这一说法？

2. 所提供的水电报告复印件是否经过认证或正式公证，证明其真实有效？

3. 当时在建筑工地是否有其他可确认的证人可以证实叶俳句和其他被告确实在建筑工地工作？

4. 报告中有没有提到具体的工作内容和完成时间？这些日期与案件中提到的合同期限一致吗？

5. 关于叶俳句手写的水电报告，能否提供笔迹样本确认笔迹一致？

6. 报告是否详细描述了每个人的工作职责和工作内容？它与案件事实的描述一致吗？

通过上述质证问题，可以对手写水电报告的真实性进行质疑和核查，确定其作为证据的有效性。同时，还需要确保其他相关证据的一致性，如工期、工种、投诉内容等的一致性。

(Regarding the cross-examination of the copy of Ye Haiku's handwritten hydropower report provided by Ye Haiqun, I suggest raising the following questions for questioning:

1. Regarding Ye Haiku's handwritten water and electricity report, are there any formal documents in the form of work records, work logs or contracts to support this statement?

2. Has the copy of the water and electricity report provided been certified or officially notarized to prove that it is true and valid?

3. Were there any other identifiable witnesses at the construction site at that time who could confirm that Ye Haiku and other respondents were actually working on the construction site?

4. Does the report mention the specific work content and completion date? Do these dates coincide with the contract periods mentioned in the case?

5. Regarding Ye Haiku's handwritten water and electricity report, can you provide a handwriting sample to confirm that the handwriting is the same?

6. Does the report describe in detail each person's job responsibilities and work content? Is it consistent with the description of the facts of the case?

Through the above cross-examination questions, the authenticity of the handwritten water and electricity report can be questioned and checked to determine its validity as evidence. At the same time, it is also necessary to ensure the consistency of other relevant evidence, such as the consistency of the construction period, type of work, and content of the complaint.)

Table 31: Example dialogue for Cross-examination Strategy Generation(6-2).

<p>Information: 全国人大常委会法制工作委员会经济法办公室主任岳仲明在回答记者提问时表示：“为确保畜禽产品有效供给和质量安全，必须大力提高畜禽养殖生产能力和水平。”此次修订明确，国家建立和完善现代畜禽养殖体系，鼓励和支持规模化、标准化、智能化养殖发展，促进种养一体化、农牧循环化、绿色发展，在畜禽繁育管理、养殖生产、屠宰管理、畜禽交易等方面提供服务。这方面强调必须遵守有关标准。</p> <p>(Yue Zhongming, director of the Economic Law Office of the Legal Affairs Committee of the Standing Committee of the National People's Congress, said in response to a reporter's question: "To ensure the effective supply and quality safety of livestock and poultry products, we must vigorously improve the production capacity and level of livestock and poultry breeding." This revision of the law clarifies that the country has established and improved The modern livestock and poultry breeding system encourages and supports the development of large-scale, standardized and intelligent breeding, promotes the integration of planting and breeding, agricultural and animal husbandry cycles, and green development, and provides services in breeding livestock and poultry management, breeding production, slaughter management, livestock and poultry trading, etc. The aspect emphasized the need to comply with relevant standards.)</p>	
<p>Needs: 给出新修订的《牧业法》的主要内容。(Give the main contents of the newly revised Animal Husbandry Law.)</p>	
<p>User: 新修订的《牧业法》明确了哪些主要内容？(What are the main contents clearly stated in the newly revised Animal Husbandry Law?)</p>	
<p>AI assistant: 新修订的《牧业法》主要提出以下内容：</p> <ol style="list-style-type: none"> 1. 明确建立健全现代畜禽养殖体系。这体现了国家对现代化、规模化、标准化和智能化的重视。 2. 鼓励和支持规模化、标准化、智能化养殖发展。这意味着国家将采取政策和其他措施，促进畜牧业从传统的小规模和分散状态转变。向大型化、集中化、高效化方向发展。 3. 推进种养融合和农牧业循环，促进绿色发展。这说明了重视环境保护，通过促进种养一体化和农牧循环，最大限度地利用资源和保护环境。 4. 强调在养殖畜禽管理、养殖生产、屠宰管理、畜禽交易等方面遵守相关标准。这表明国家对畜禽产品有效供给和质量安全的重视。通过规范管理和制定标准，努力确保畜禽产品质量安全。 <p>(The newly revised Animal Husbandry Law mainly proposes the following contents:</p> <ol style="list-style-type: none"> 1. Make it clear that the country will establish and improve a modern livestock and poultry breeding system. This represents the country's emphasis on modernization, scale, standardization and intelligent breeding. 2. Encourage and support the development of large-scale, standardized and intelligent breeding. This means that the state will adopt policies and other measures to promote the transformation of animal husbandry from the traditional small-scale and scattered state. Develop towards large-scale, centralized and efficient development. 3. Promote the integration of planting and breeding and the cycle of agriculture and animal husbandry, and promote green development. This illustrates the emphasis on environmental protection and maximizing the utilization of resources and protecting the environment by promoting the integration of planting and breeding and the cycle of agriculture and animal husbandry. 4. Emphasize compliance with relevant standards in aspects such as breeding livestock and poultry management, breeding production, slaughter management, livestock and poultry trading, etc. This shows that the country attaches great importance to the effective supply and quality safety of livestock and poultry products. Strive to ensure the quality and safety of livestock and poultry products through standardized management and formulation of standards.) 	

Table 32: Example dialogue for Litigation Strategy Generation(6-3).

Information: 经审理查明: 2018年8月13日10时许, 被告人A在某县某镇某道路上经过某幼儿园时, 看到受害人b1 (2014年11月3日出生) 独自玩耍。出于邪念, 他偷偷地把奕一带走了。他的母亲发现孩子失踪后报了警。当日13时许, A在某市某区某工厂附近被俘。年幼的孩子B1被他的父母救出并带了回来。被告人A在开庭时对上述事实无异议, 并有案件登记表、出庭记录、户籍信息、现场监控、出生证明、犯罪记录核实证明、刑事判决及释放证明、证人C、丁某、吴某的证言及被告人A在公安机关的供词等证据充分证实。

(After trial, it was found that: At about 10 o'clock on August 13, 2018, defendant A passed by a kindergarten on a certain road in a certain town in a certain county and saw the victim B 1 (born on November 3, 2014) playing alone. Out of evil thoughts, he secretly took Yi 1 away. His mother called the police after discovering that the child was missing. At about 13:00 that day, A was captured near a factory in a certain district of a certain city. The young child B1 was rescued and taken back by his parents. Defendant A had no objection to the above facts during the court session, and had a case registration form, attendance at the case, household registration information, on-site surveillance, birth certificate, criminal record verification certificate, criminal judgment and release certificate, witnesses C and Ding , Wu's testimony and defendant A's confession in the public security organ and other evidence are sufficient to confirm.)

Needs: 给出判决结果。(Give the judgement result.)

User: 请根据以上案件事实帮我写一份判决结果部分。(Please help me write a judgment result part based on the above case facts.)

AI assistant: 本院认为, 被告人A以非法手段剥夺他人人身自由, 其行为构成拐骗儿童罪。被告人A在盗窃儿童后被公安机关直接抓获, 随后未对被害人B1造成其他身体或精神伤害。因此, 应当依法从轻处罚被告人A。根据《中华人民共和国刑法》第一百四十一条规定, 犯本罪的, 处五年以下有期徒刑或者拘役。考虑到被告A的行为情节及对社会的危害程度, 判决如下: 被告A犯拐骗儿童罪, 判处有期徒刑X年。(“X”根据具体情况的严重程度填写。) 以上各项必须遵守。当事人有异议的, 可以自判决书送达之日起十日内, 通过本人或者指定的辩护人向本院提出上诉。

(This court held that defendant A used illegal means to deprive others of their personal freedom, and his behavior constituted the crime of child abduction. Defendant A was directly captured by the public security organs after stealing the child, and subsequently did not cause any other physical or mental harm to victim B1. Therefore, defendant A should be given a lighter sentence in accordance with the law. According to Article 141 of the Criminal Law of the People's Republic of China, anyone who commits this crime shall be sentenced to fixed-term imprisonment of not more than five years or criminal detention. Taking into account the circumstances of defendant A's behavior and the degree of harm to society, the ruling is as follows: defendant A is guilty of child abduction and sentenced to X years in prison. (The "X" is filled in according to the severity of the specific situation.) The above must be complied with. If you have any objections, you can appeal to this court through yourself or the appointed defender within ten days from the date of delivery of the judgment.)

Table 33: Example dialogue for Judgment Document Generation(6-4).

Information: 经依法调查查明, 被告人卢某某于2022年8月23日13时许, 因与被害人闫某某发生事端, 前往北京市怀柔区**镇** ** ** ** **楼** **单元。在争吵中, 被告卢某用剪刀刺伤了被害人闫某。经鉴定, 受害人严某某的身体伤害程度为轻伤。被告人卢某某案发后主动向公安机关投案自首, 并于事后如实供述上述事实。犯罪工具已被缴获。确定上述事实的证据如下: 被告人供述与辩护: 陆被告人供述与辩护; 2. 受害人陈述: 受害人严某陈述; 3. 证人证言: 证人高、史证言; 4. 鉴定意见: 北京市怀柔区公安法医鉴定中心鉴定证书、北京市公安法医鉴定中心鉴定证书; 5. 证明文件: 查封决定书、查封笔录、查封清单、门诊病历、诊断证明等; 6. 物证: 犯罪现场照片、涉案物品照片、犯罪工具照片; 7. 视听资料: 警察执法录像; 8. 其他证明材料: 110警察收据、案件登记表、案件到达史、户籍证明、工作记录、过往犯罪记录材料等。上述证据收集程序合法, 内容客观真实, 足以认定指控事实。被告人陆某某对所指控的犯罪事实和证据没有异议, 主动认罪接受处罚。

(After investigation in accordance with the law, it was found that the defendant Lu Moumou went to Unit **, Building ** Home **, ** Town, Huairou District, Beijing at about 13:00 on August 23, 2022, because of an incident with the victim Yan Moumou. During the quarrel, the defendant Lu stabbed the victim Yan with scissors. After identification, the degree of physical injury suffered by the victim Yan Moumou was minor injury. The defendant Lu Moumou took the initiative to surrender to the public security organ after the incident, and later truthfully confessed the above facts. The tools of the crime have been seized. The evidence for determining the above facts is as follows: 1. Defendant's confession and defense: defendant Lu's confession and defense; 2. Victim's statement: victim Yan's statement; 3. Witness testimony: witnesses Gao and Shi Testimony; 4. Appraisal opinion: Appraisal certificate from Beijing Huairou District Public Security Forensic Appraisal Center, Beijing Public Security Forensic Appraisal Center appraisal certificate; 5. Documentary evidence: seizure decision, seizure transcript, seizure list, outpatient medical records, diagnosis certificate, etc.; 6. Physical evidence: pictures of the place where the crime occurred, pictures of items involved, and photos of crime tools; 7. Audio-visual materials: police law enforcement videos; 8. Other supporting materials: 110 police receipt, case registration form, case arrival history, household registration certificate, work records, Materials on previous criminal records, etc. The above-mentioned evidence collection procedures are legal and the content is objective and true, which is sufficient to determine the facts of the accusation. The defendant Lu Moumou had no objection to the alleged criminal facts and evidence, and voluntarily pleaded guilty and accepted punishment.)

Needs: 撰写起诉书的结论部分。(Writing the conclusion part of the indictment.)

User: 请给出起诉书的结论部分。(Please write the conclusion of the indictment.)

AI assistant: 本院认为, 被告人陆某某的行为构成故意伤害罪。被告人卢某在与被害人闫某发生口角后, 选择使用暴力手段, 用剪刀将被害人闫某刺伤, 致使被害人闫某受伤。这严重侵犯了他人的人身权利, 对社会有害。威胁公共秩序。被害人闫某的伤情被认定为一级轻伤, 这清楚地表明被告人吕某的行为确实对被害人闫某造成了实际伤害。

鉴于被告人陆某某事后主动自首, 并如实供述所犯罪行, 表明其有悔过之心。结合上述证据, 真实、全面地反映了案件事实。因此, 在量刑时应适当考虑从轻或减刑。

因此, 根据《中华人民共和国刑法》第二百三十四条的规定, 建议判处被告人陆某某故意伤害罪。量刑时应考虑其自首和如实供述的情节, 从轻或从轻量刑。减轻处罚。兹提起诉讼, 请求依法判决。(In the opinion of the court, the defendant Lu Moumou's behavior constituted the crime of intentional injury. After a quarrel with the victim Yan, the defendant Lu chose to use violent means and stabbed the victim Yan with scissors, causing the victim Yan to be injured. This has seriously violated the personal rights of others and is harmful to society. Threat to public order. The injury of the victim Yan was assessed as a minor injury of the first degree, which clearly shows that the behavior of the defendant Lu actually caused actual harm to the victim Yan.

In view of the fact that the defendant Lu Moumou voluntarily surrendered after the incident and truthfully confessed the crime he committed, it shows that he has remorse. Combined with the above evidence, it truly and comprehensively reflects the facts of the case. Therefore, appropriate consideration should be given to lenient or reduced punishment when sentencing.

Therefore, in accordance with the provisions of Article 234 of the Criminal Law of the People's Republic of China, it is recommended that the defendant Lu Moumou be convicted of the crime of intentional injury. When sentencing, the circumstances of his voluntary surrender and truthful confession should be taken into consideration, and he should be given a lighter sentence or a lighter sentence. Reduce penalties. We hereby sue and request to be sentenced according to law.)

Table 34: Example dialogue for Procuratorial Document Generation(6-5).

Information: 房子位于深圳龙岗区，价值200万。这辆车是2020年购买的宝马320，车牌号为广东B123456，价值30万美元。存款在中国银行。账号是43201234567891。(The house is in Longge District, Shenzhen, worth 2 million. The car is a BMW 320 purchased in 2020 with license plate number Guangdong B123456, worth 300,000. The deposit is in the Bank of China. The account number is 43201234567891.)

Needs: 写一份遗嘱。(Write a will.)

User: 你好，AI助手。我需要你帮我写遗嘱。我拥有一些房产，包括：位于深圳龙岗的一套房产，价值200万元；2020年购买的宝马320，车牌号广东B123456，价值30万元；在中国的银行账户存款为43201234567891。你能帮我起草一份合适的遗嘱吗？

(Hello, AI assistant. I need your help writing a will. I own some properties, including: a property located in Longge District, Shenzhen, worth 2 million yuan; a BMW 320 car purchased in 2020, with the license plate number Guangdong B123456, worth 300,000 yuan; and a bank account in China Deposit of 43201234567891. Can you help me draft a suitable will?)

AI assistant: 我，[你的全名]，在完全理智的状态下，凭我自己的自由意志立下这个遗嘱。

1. 财产

这将包括我名下的财产，具体内容如下：

1. 我名下的房产位于深圳市龙岗区，价值约200万元。
2. 这辆2020年购买的宝马320车牌号为广东B123456，价值约30万元。
3. 存款在中国银行账户43201234567891。

2. 分配

我已决定在我死后，上述所有个人财产将按以下方式处理：

[在本节中，您需要列出每个遗产的受益人及其获得的份额]

3. 遗嘱执行人

本人现委任【遗嘱执行人全名及地址】为本人遗嘱执行人。

4. 法律效力

本遗嘱表达了我的最终和无条件的决定，无论我在任何其他时间可能做出的口头或书面的任何不同或矛盾的陈述。

本遗嘱自本人签名之日起立即生效。我明白，如果我不签字并注明日期，本遗嘱将不生效。

[你的全名]

I, [your full name], make this will of my own free will and in a perfectly sensible state of mind.

1. Property

This will covers the property in my name and is detailed as follows:

1. A property under my name located in Longgang District, Shenzhen City is worth about 2 million yuan.
2. The BMW 320 car purchased in 2020 has the license plate number Guangdong B123456 and is worth approximately 300,000 yuan.
3. Deposits deposited in Bank of China account 43201234567891.

2. Distribution

I have decided that all of my personal property mentioned above will be disposed of in the following manner after my death:

[In this section you will need to list the beneficiaries of each estate and the share they received]

3. Executor

I hereby appoint [executor's full name and address] as executor of my will.

4. Legal effect

If any part of this Will is declared invalid by law, the remaining parts shall remain in full force and effect. This will expresses my final and unconditional decision, regardless of any different or contradictory statements, whether oral or written, that I may make at any other time.

This will is effective immediately upon my personal signature and date. I understand that this will will not be effective if I do not sign and date it. [your full name]

Table 35: Example dialogue for Lawyer Document Generation(6-6).

Information: 2019年1月至2020年8月，被告人王某某与同案犯岳某某、李某某等人，在靖阳县XX镇XX村以南的荒地上，采用了爆破、铲土等方法。盗墓9座，鎏金青铜龙头带钩、唐代鸟纹菱形镜等文物被盗，卖给杨某某等人。两件涉案文物被追回。经鉴定，此次发掘对该古墓的历史科学价值造成了一定程度的破坏，也对墓区的完整性及其历史科学价值造成了一定程度的破坏。被盗的战国至汉代的皮带钩被列为三级文物。唐代鸟纹菱形镜是一种普遍的文物。(From January 2019 to August 2020, the defendant Wang **, together with his accomplices Yue Moumou, Li Moumou, and others, used methods such as blasting holes and shoveling in the wasteland south of XX Village, XX Town, Jingyang County. Nine ancient tombs were robbed, and cultural relics such as a gilt bronze dragon head belt hook, a Tang Dynasty bird-patterned rhombus mirror, etc. were stolen and sold to Yang Moumou and others. Two cultural relics involved in the case were recovered. It has been identified that the excavation has caused a certain degree of damage to the historical and scientific value of the ancient tomb, and has also caused a certain degree of damage to the integrity of the tomb area and its historical and scientific value. The stolen belt hooks from the Warring States Period to the Han Dynasty are classified as third-level cultural relics. The Tang Dynasty bird-patterned rhombus mirror is a general cultural relic.)

Needs: (确定与案件相关的法律条款。)Determine the legal provisions relevant to the case.

User: 本案涉及的主要法律条文有哪些? (What are the main legal provisions involved in this case?)

internlm-chat-7b: 《中华人民共和国文物保护法》("Cultural Relics Protection Law of the People's Republic of China")

User: 请问《中华人民共和国文物保护法》的哪些规定与本案有关? (Could you please provide us with details: Which provisions of the "Cultural Relics Protection Law of the People's Republic of China" are relevant to this case?)

internlm-chat-7b: 《中华人民共和国文物保护法》与本案有关的主要规定是:

1. 第二十六条禁止伪造、涂改受保护的文物，禁止擅自故意损坏、污损、移动、拆除受保护的文物。
2. 第三十条禁止将文物和文物保护单位及其保护范围内出土的文物买卖、抵押或者挪作他用。
3. 第三十五条禁止私自陈列、展览文物、文物。
4. 第四十一条禁止在文物保护单位内开展影响文物、文物安全的活动。
5. 第五十二条禁止在文物保护单位内进行采矿、挖掘、采石和其他可能影响文物、文物安全的活动。

The main provisions related to this case in the Cultural Relics Protection Law of the People's Republic of China are:

1. Article 26: It is prohibited to forge or alter protected cultural relics and to intentionally damage, deface, move or dismantle protected cultural relics without authorization.
 2. Article 30: It is prohibited to buy, sell, mortgage or transfer cultural relics and relics unearthed within cultural relics protection units and their protection scope for other purposes.
 3. Article 35: Private display or exhibition of cultural relics and relics is prohibited.
 4. Article 41: It is prohibited to carry out activities that affect the safety of cultural relics and relics in cultural relics protection units.
 5. Article 52: Mining, excavation, quarrying and other activities that may affect the safety of cultural relics and relics are prohibited within cultural relics protection units.
-

Table 36: Example dialogue that gives short responses. LLM did not give specific legal terms at the beginning until the user gave clearer instructions.

<p>Information: 公诉机关指控：2015年底至2016年5月9日，被告人刘某分三次向黄某、周某出售冰毒2.5克（俗称“冰”），并于2016年5月收取毒资900元。3月12日，被告人刘某在准备向黄某出售冰毒时被抓获，从其身上搜获冰毒0.7720克、电子秤1个。针对上述指控的事实，检察机关向法院提交了相关证据，予以证实。据此，公诉机关认为，被告人刘某的行为违反了《中华人民共和国刑法》xxx、xx的规定，应当以贩毒罪追究其刑事责任。请求本院依法处罚。</p> <p>The public prosecution alleges: From the end of 2015 to May 9, 2016, the defendant Liu sold 2.5 grams of methamphetamine (commonly known as "ice") to Huang and Zhou on three occasions, and collected drug money of 900 yuan in May 2016. On March 12, the defendant Liu was caught when he was preparing to sell methamphetamine to Huang A, and 0.7720 grams of methamphetamine and an electronic scale were seized from him. In response to the facts of the above accusations, the public prosecution agency presented relevant evidence to the court to substantiate them. Based on this, the public prosecutor's office believed that defendant Liu's behavior had violated the provisions of xxx and xx in the Criminal Law of the People's Republic of China, and he should be held criminally responsible for the crime of drug trafficking. Request this court to punish according to law.</p>
<p>Needs: 确定被告所要求的赔偿金额。(Determine the amount of compensation required by the defendant.)</p>
<p>User: 根据公诉机关的指控，被告人刘某因贩毒可能需要支付多少赔偿？(According to the charges brought by the public prosecution, how much compensation may the defendant Liu have to pay for drug trafficking?)</p> <p>AI assistant: 这个问题需要根据具体的法律规定和案件事实来判断。一般来说，根据《中华人民共和国刑法》的规定，贩毒活动应当承担相应的刑事责任和赔偿义务。赔偿数额需要根据案件的具体情况确定，包括被告销售毒品的数量、违法所得的程度、对被害人的影响等因素。因此，具体的赔偿金额需要根据具体情况进行预测。</p> <p>This issue needs to be judged based on specific legal provisions and case facts. Generally speaking, according to the provisions of the "Criminal Law of the People's Republic of China", drug trafficking activities should bear corresponding criminal liability and compensation obligations. The amount of compensation needs to be determined based on the specific circumstances of the case, including the quantity of drugs sold by the defendant, the extent of illegal gains, the impact on the victim and other factors. Therefore, the specific amount of compensation needs to be predicted based on the specific circumstances.</p> <p>User: 我理解你的解释，但在本案中，被告人刘某共出售冰毒3.272克，收取毒款900元。考虑到这些具体的案例因素，你能帮我预测一下刘先生可能需要支付的赔偿金额吗？</p> <p>I understand your explanation, but in this case, the defendant Liu sold a total of 3.272 grams of methamphetamine and received 900 yuan in drug money. Taking into account these specific case factors, can you help me predict the amount of compensation Liu may need to pay?</p> <p>AI assistant: 根据你提供的案件情况，被告人刘某出售冰毒3.272克，收取毒款900元，需要支付的赔偿金额可能较高。具体价值需要根据案件的具体情况来判断，如刘某是否赔偿了受害人的损失，是否有其他的赔偿方式等。</p> <p>According to the case situation you provided, the defendant Liu sold 3.272 grams of methamphetamine and received 900 yuan in drug money, so the amount of compensation he may need to pay may be relatively high. The specific value needs to be judged based on the specific circumstances of the case, such as whether Liu has compensated the victim for his losses, and whether there are other ways of compensation.</p> <p>User: 在这种情况下，我理解我们无法达成确切的赔偿金额。但一般来说，在这类毒品交易案件中，赔偿金额的大致范围是多少？</p> <p>In this case, I understand that we cannot arrive at an exact amount of compensation. But generally speaking, what is the approximate range of the amount of compensation required to be paid in such drug dealing cases?</p> <p>AI assistant: 在这种情况下，需要支付的赔偿金额可能从几千元到几万元不等。具体价值需要根据案件的具体情况来判断，如刘某是否赔偿了受害人的损失，是否有其他的赔偿方式等。</p> <p>In this case, the amount of compensation required to be paid may range from a few thousand yuan to tens of thousands of yuan. The specific value needs to be judged based on the specific circumstances of the case, such as whether Liu has compensated the victim for his losses, and whether there are other ways of compensation.</p>

Table 37: Example dialogue that doesn't give helpful answers. Despite adequate prompts from the user, DISC-LawLLM still refused to calculate the defendant's compensation amount.