LawBench: Benchmarking Legal Knowledge of Large Language Models

Zhiwei Fei[1*], **Xiaoyu Shen**[2*], **Dawei Zhu**[3*], **Fengzhe Zhou**[4], **Zhuo Han**[1], **Alan Huang**[5], **Songyang Zhang**[4], **Kai Chen**[4], **Zhixin Yin**[1], **Zongwen Shen**[1], **Jidong Ge**[1†], **Vincent Ng**[6]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Digital Twin Institute, Eastern Institute of Technology, Ningbo [3]Saarland University
[4]Shanghai AI Laboratory, China [5]School of Science and Engineering Magnet, USA
[6]University of Texas at Dallas, USA
602022320001@smail.nju.edu.cn, gjd@nju.edu.cn

Abstract

We present LawBench, the first evaluation benchmark composed of 20 tasks aimed to assess the ability of large language models (LLMs) to perform Chinese legal-related tasks. LawBench is meticulously crafted to enable precise assessment of LLMs' legal capabilities from three cognitive levels that correspond to the widely accepted Bloom's cognitive taxonomy. Using LawBench, we present a comprehensive evaluation of 21 popular LLMs and the first comparative analysis of the empirical results in order to reveal their relative strengths and weaknesses. All data, model predictions and evaluation code are accessible from https://github.com/open-compass/LawBench.

1 Introduction

Legal tasks encompass a broad spectrum of applications, predominantly text-based, necessitating comprehension and interpretation of highly professional legal text. Currently, they are primarily conducted by legal experts, who require years of extensive specialized training to process legal cases. Endowing large language models (LLMs) with legal expertise can not only improve the working efficiency of legal officers, but also address the overwhelming demand of legal assistance from non-professionals, thereby improving public access to justice (Cui et al., 2022; Trozze et al., 2024).

There have been works assembling legal-related tasks, including LexGLUE (Chalkidis et al., 2022), LBOX OPEN (Hwang et al., 2022), LEXTREME (Niklaus et al., 2023), SCALE (Rasiah et al., 2023), and LegalBench (Guha et al., 2024). LegalBench, in particular, presented the first steps towards constructing an interdisciplinary collaborative legal reasoning benchmark for the English language and evaluated 20 LLMs in 162 legal tasks. While these benchmarks have been developed for legal systems that span a range of countries and jurisdictions, it is worth noting that these countries and jurisdictions have all adopted the common law system.

Unlike the common law system, which is widely accepted in the western countries, the Chinese legal system is rooted in the civil law family. Judges in the civil law system are obliged to respect the established statutory law articles and ground their decisions on them. Understanding and applying existing statutes and codes, rather than studies of precedents, are of paramount importance (Zheng, 1986). Hence, models that perform well in the aforementioned benchmarks may not necessarily perform well on the Chinese legal tasks, and it is necessary to develop a benchmark that emphasizes the required skill set for the Chinese legal system.

With this in mind, we present *LawBench*, the first evaluation benchmark composed of 20 tasks aimed to assess LLMs' capabilities on performing legal-related tasks under the Chinese civil law system. Alongside LawBench, our contributions in this paper are three-fold. First, we develop a taxonomy that divides these 20 tasks into three skill levels according to widely accepted Bloom's cognitive taxonomy (Krathwohl, 2002), providing suggestive evidence that LawBench is capable of assessing an LLM's ability to *memorize*, *understand*, and *apply* legal knowledge. Second, we perform an extensive evaluation of 21 popular LLMs on LawBench to assess their ability to perform Chinese legal tasks. Third, we conduct the first comparative analysis of these results in order to gain insights into the relative strengths and weaknesses of the LLMs. Finally, to stimulate work in this area of research, we integrate the benchmark and evaluation code into the OpenCompass platform (Contributors, 2023).

2 Related Work

In this section, we focus our discussion on existing benchmarks developed for the legal domain.

[*]Equal Contribution, [†]Corresponding author.

Chinese benchmarks. Our benchmark, LawBench, is the first evaluation benchmark developed for the Chinese legal domain. LawBench has influenced the design of a subsequently released benchmark, LAiW (Dai et al., 2023), according to its authors. Among the 14 tasks in LAiW, six can be found in LawBench, three are similar to a task in LawBench, and five are not present in LawBench.[1]

A few points deserve mention. First, unlike LawBench, LAiW does not contain any *legal knowledge memorization* tasks (see Section 3.1). Second, there are a number of critical legal application tasks that are present in LawBench but not LAiW, including tasks that involve proofreading a legal document as well as tasks that involve predicting an amount (e.g., the monetary penalty) associated with a case. We believe that the failure to include such application tasks will disallow LAiW to evaluate a model in certain important scenarios.

DISC-Law-Eval (Yue et al., 2023) is another Chinese legal benchmark that is primarily focused on assessing a model's capabilities on legal exams, with an emphasis on multiple-choice questions. In addition evaluating a model's performance on multiple-choice questions, LawBench examines a model's proficiency in other tasks such as information extraction and reading comprehension, and ensures that the scope of testing is not confined to skills required for judicial examinations.

Non-Chinese benchmarks. Other prominent legal benchmarks include LexGLUE (Chalkidis et al., 2022), which focuses on EU and American laws, LBOX OPEN (Hwang et al., 2022), which focuses on South Korean laws, LegalBench (Guha et al., 2024), which focuses on American laws, LEXTREME (Niklaus et al., 2023), which covers 24 Indo-European and Uralic languages, and SCALE (Rasiah et al., 2023), which covers three Swiss languages across a range of legal tasks. These benchmarks are designed to evaluate the legal logic of a model when used with common law jurisdictions, which differs significantly from civil law jurisdictions, such as the one adopted in China.[2] As a result, they have distinct application scenarios and processes. In Chinese judicial practice, tasks such as document review and dispute focus detection are crucial but are not present in these benchmarks. Given the uniqueness of the Chinese legal system and the diversity of the tasks in LawBench, we believe LawBench will contribute to the multilinguality of global legal research.

We conclude this section by noting that in the aforementioned papers, model evaluations on the benchmarks are often accompanied by shallow or even no analysis of the results. For instance, we consider the analysis in the LegalBench paper shallow because it was solely performed based on the average performance of the models over all tasks. In contrast, a key contribution of our work is a comparative analysis of the models not only w.r.t. their average performance but also at the task level.

3 LawBench Construction

In this section, we provide a detailed description of the principles behind the design of LawBench.

3.1 The Hierarchical Ability Taxonomy

Rather than categorize tasks based on their difficulty (Huang et al., 2024), we propose to employ the widely used Bloom's cognitive model (Krathwohl, 2002) to classify tasks into different dimensions (Yu et al., 2024). Bloom's Taxonomy system was initially proposed by the educational psychologist Benjamin Bloom and his collaborators in 1956 and has been widely applied since then. In particular, it has effectively aided teachers in curriculum design and the assessment of student learning outcomes. Bloom's Taxonomy divides learning objectives in the cognitive domain into six levels, from the lowest to the highest: Remember, Understand, Apply, Analyze, Evaluate, and Create. These levels describe the depth and complexity of cognitive learning and provide an organized framework. Teachers can use Bloom's Taxonomy to ensure diversity and completeness in course objectives. By combining the learning objectives at different levels, one can promote comprehensive student development, encouraging them to progress from simple memorization and understanding to higher-level analysis, evaluation, and creation.

Inspired by this classification approach, we simplified Bloom's cognitive hierarchy model and kept the first three categories in Bloom's taxonomy to assess the legal knowledge of LLMs[3]:

1. Knowledge Memorization: The memorization level measures the basic requirement of remembering legal-related knowledge. It tests LLMs' ability in the memorization and recitation of basic

[1] A description of the task similarities and differences between LawBench and LAiW can be found in Appendix A.

[2] See Appendix B for a discussion of these jurisdictions.

[3] See Appendix C for more discussion on using Bloom's Taxonomy to assess LLMs' capabilities in legal scenarios.

Cognitive Level	ID	Task	Data Source	Metric	Type
Legal Knowledge Memorization	1-1	Article Recitation	FLK	Rouge-L	Generation
	1-2	Knowledge Question Answering	JEC_QA	Accuracy	SLC
Legal Knowledge Understanding	2-1	Document Proofreading	CAIL2022	F0.5	Generation
	2-2	Dispute Focus Identification	LAIC2021	F1	SLC
	2-3	Marital Disputes Identification	AIStudio	F1	MLC
	2-4	Issue Topic Identification	CrimeKgAssitant	Accuracy	SLC
	2-5	Reading Comprehension	CAIL2019	rc-F1	Extraction
	2-6	Named-Entity Recognition	CAIL2022	soft-F1	Extraction
	2-7	Opinion Summarization	CAIL2021	Rouge-L	Generation
	2-8	Argument Mining	CAIL2022	Accuracy	SLC
	2-9	Event Detection	LEVEN	F1	MLC
	2-10	Trigger Word Extraction	LEVEN	soft-F1	Extraction
Legal Knowledge Application	3-1	Fact-based Article Prediction	CAIL2018	F1	MLC
	3-2	Scene-based Article Prediction	LawGPT	Rouge-L	Generation
	3-3	Charge Prediction	CAIL2018	F1	MLC
	3-4	Prison Term Prediction w.o. Article	CAIL2018	nLog-distance	Regression
	3-5	Prison Term Prediction w. Article	CAIL2018	nLog-distance	Regression
	3-6	Case Analysis	JEC_QA	Accuracy	SLC
	3-7	Criminal Damages Calculation	LAIC2021	Accuracy	Regression
	3-8	Consultation	hualv.com	Rouge-L	Generation

Table 1: Task list for LawBench.

legal knowledge such as regulations, cases, concepts, common sense, legal facts and terminologies.

2. Knowledge Understanding: The understanding level involves understanding the meanings and connotations of legal documents. This includes the ability to comprehend and interpret legal concepts, text, and issues, such as identifying the entities and relationships within legal texts, and detecting types of legal issues and points of dispute.

3. Knowledge Application: The application level requires LLMs to integrate legal knowledge, reason over it and address real-world legal cases. It evaluates a model's logical reasoning abilities to perform legal consultation, judicial assistance, as well its as numerical reasoning abilities.

3.2 Data Sources and Selected Tasks

We selected 20 tasks falling under the aforementioned capability levels.[4] The task list is shown in Table 1. For easy reference, every task is assigned a unique task id. Below we describe how the tasks are divided into the three cognitive levels.

Legal knowledge memorization tasks examine the extent to which LLMs encode legal knowledge within their parameters. There are two major types of legal knowledge that require memorization: (1) core law articles and regulation content and (2) other fundamental legal concepts, notions and rules. We construct two tasks corresponding to these two types of knowledge (1-1 and 1-2).

Legal knowledge understanding tasks examine the extent to which LLMs can comprehend the entities, events, and relationships within legal text. Understanding legal text is a pre-condition to utilizing the knowledge in concrete downstream applications (Cui et al., 2022). We selected 10 tasks for this cognitive level (2-1 through 2-10).

Legal knowledge application tasks examine the ability of LLMs to not only understand legal knowledge but also simulate law professionals to apply the knowledge in solving realistic legal tasks. In the task design, we extensively examined a model's different reasoning abilities via three legal content reasoning tasks (legal judgement prediction (3-1 through 3-5), case analysis (3-6), and consultation (3-8)) and one numerical reasoning task (criminal damages calculation (3-7)).

When predicting case judgments, judges follow a certain order when hearing a case (Zhong et al., 2018; Huang et al., 2021). Therefore, when constructing the legal judgment prediction task, we simulated this process by decomposing the CAIL2018 dataset into three tasks: fact-based article prediction (3-1), charge prediction (3-3) and prison term prediction. We further consider the task of prison term prediction in two scenarios, one without using article content (3-4) and one using article content (3-5) to examine LLMs' capability in utilizing the article content to make accurate judgement predictions. Besides, we also add the task scene-based fact prediction to simulate judges' recognition of legal provisions (3-2).

[4] Some tasks may belong to more than one category. We have categorized them based on their primary capabilities.

ID	Definition	Data Collection Details
1-1	Article Recitation: Given a law article number, recite the article content.	We collected the contents of laws and regulations from the national database (see Appedix F FLK part) and consulted students with a legal background to select 152 sub-laws under the 5 core laws. We further incorporated updated laws and regulations, including constitutional amendments, to evaluate the model's ability to comprehend legal changes.
1-2	Knowledge Question Answering: Given a question asking about basic legal knowledge, select the correct answer from 4 candidates.	We collect knowledge-based questions from the JEC-QA tasks (Zhong et al., 2020). To simplify the process of locating answers during the test, we exclusively chose single-label questions from them.

Table 2: Legal knowledge memorization tasks: definition and data collection details.

As can be seen in Table 1, the 20 tasks can be divided into five task types: generation, single-label classification (SLC), multi-label classification (MLC), regression, and extraction. The table also shows for each task the evaluation metric and the source from which we sampled the test instances.[5]

Two points deserve mention. First, for each task, we crafted a set of 500 test instances. This decision was motivated by LegalBench, taking into account the time required to validate a model's performance. Second, while for the majority of tasks the test instances were sampled from existing sources, those for Tasks 1-1 and 3-8 were collected by us.

A detailed description of these tasks can be found in Tables 2, 3 and 4.

4 Evaluation

4.1 Experimental Setup

Evaluation settings. We evaluate the 21 LLMs on each task in LawBench via prompting in both zero-shot and one-shot settings. For zero-shot inference, the model input is merely the task instruction and the query.[6] To build the model input for one-shot inference, a single example randomly sampled from outside the test set, which includes the query and the corresponding answer, is added to the input for zero-shot inference, between the instruction and the actual query to the model.

Answer extraction. Since most LLMs are generative models, when using them to accomplish a task, it is necessary to extract the answers from the generated content before they can be automatically evaluated (Adlakha et al., 2023). Details on answer extraction can be found in Appendix G.

Evaluation metrics. After answer extraction, we evaluate answer quality using automatic evaluation metrics. The metric that we use to evaluate each task is shown in the "Metric" column in Table 1. As can be seen, the SLC, MLC, and extraction tasks are evaluated using variants of F-measure, including F0.5, F1, rc-F1, and soft-F1. The regression tasks are evaluated using nLog-distance. The generation tasks are evaluated using Rouge-L, except for Document Proofreading (2-1), which is evaluated using F0.5. All metric values are between 0 and 100, with higher values indicating better outputs.[7]

A point about the choice of metrics deserves mention. Except for Consultation (3-8), which is a task we defined, for each of the remaining tasks the metric we employed is the one officially used to evaluate the task on the corresponding data source shown in Table 1.

Inference. We employ OpenCompass (Contributors, 2023) to perform model inference. For ChatGPT and GPT-4, we set the temperature too 0.7 and the top p to 1.[8] For other chat models, we tailor the prompt using prefixes and suffixes specific to each model. Greedy decoding is performed during generation for all open-sourced models. We set the input token length limit to 2048 and an output token length to 1024.[9] Right truncation is performed for input prompts exceeding the length limitation.

Models. We evaluate a wide spectrum of LLMs of various sizes, grouping them into three major categories based on their pre-training and fine-tuning domains: multilingual LLMs, Chinese-oriented LLMs and legal-specific LLMs. Specifically, we employ (1) six multilingual LLMs, including four open-sourced LLMs and two commercial LLMs; (2) nine Chinese-oriented LLMs, which are pre-trained on Chinese text and therefore typically perform better than the multilingual models on Chi-

[5] See Appendix F for details on the data sources.

[6] See Appendix D for per-task instructions and example queries, and Appendix E for details on prompt engineering.

[7] Definitions of these metrics can be found in Appendix G.

[8] We set the temperature to 0.7 because it is the default setting. For a fairer comparison with other models, we show their results when temperature is set to 0 in Appendix I.

[9] See Appendix J for a discussion of the number of documents that exceed the maximum token length limit.

ID	Definition	Data Collection Details
2-1	Document Proofreading: Given a sentence extracted from legal documents, correct its spelling, grammar and ordering mistakes, return the corrected sentence	Legal documents, as carriers of judicial authorities and the exercise of legal rights by citizens, demand utmost precision in their textual content. We sample the original and corrected legal sentences from the CAIL2022 document proofreading task. Possible mistake types are inserted into the instructions to let the model directly output the corrected sentence.
2-2	Dispute Focus Identification: Given the original claims and responses of the plaintiff and defendant, detect the points of dispute.	In civil cases, the points of dispute represent the core of conflicts, intersection of contradictions, and issues over which the parties involved in the case are in contention. The automated recognition and detection of points of contention have practical significance and necessity for the development of the rule of law in our country. Specifically, we will provide the trial-related content from judgment documents, including the sections on claims and responses. The cases involve various legal matters such as civil loans, divorce, motor vehicle traffic accident liability, financial loan contracts, and more. We have carefully selected common types of points of contention from LAIC2021 to construct this test set.
2-3	Marital Disputes Identification: Given a sentence describing marital disputes, classify it into one of the 20 pre-defined dispute types.	Marital disputes refer to the total sum of various disputes arising from love, marriage, and divorce. Among civil disputes, marital disputes are a common type of dispute. We have selected a publicly available marriage text classification dataset on AiStudio(see Appendix F AiStudio part). This dataset consists of 20 categories, and a single text entry may have multiple labels.
2-4	Issue Topic Identification: Given a user inquiry, assign it into one of pre-defined topics.	User inquiries are typically vague. Identifying the relevant topics in legal consultations can help legal professionals better pinpoint key issues. We obtain the data from the CrimeKgAssistant project(see Appendix F CrimeKgAssistant part). We keep the most frequent 20 classes and sample 25 questions for each class to form our final test set.
2-5	Reading Comprehension: Given a judgement document and a corresponding question, extract relevant content from it to answer the question.	Judicial documents contain rich case information, such as time, location, and character relationships. Intelligently reading and comprehending judicial documents through large language models can assist judges, lawyers, and the general public in obtaining the necessary information quickly and conveniently. We use the CAIL2019 reading comprehension dataset to build this task, removing question types related to binary and unanswerable questions. We retain single and multiple-segment data as our test set.
2-6	Named-Entity Recognition: Given a sentence from a judgement document, extract entity information corresponding to a set of pre-defined entity types such as suspect, victim or evidence.	We sampled 500 examples from the CAIL2022 Information Extraction dataset as our test set. These 500 samples contain 10 entity types related to theft crimes.
2-7	Opinion Summarization: Given a legal-related public news report, generate a concise summary.	Legal summaries typically include key facts of the case, points of contention, legal issues, legal principles applied, and the judgment's outcome. It can provide a quick overview of the case content to improve the efficiency of legal professionals.
2-8	Argument Mining: Given a plaintiff's perspective and five candidate defendant's viewpoints, select one viewpoint that can form a point of dispute with the plaintiff's perspective.	In court's trial process, judgment documents play a crucial role in recording the arguments and evidence presented by both the plaintiff and the defendant. Due to differences in their positions and perspectives, as well as inconsistencies in their factual statements, disputes arise between the plaintiff and the defendant during the trial process. These points of contention are the key to the entire trial and the essence of judgment documents. This task aims to extract valuable arguments and supporting materials from a large volume of legal texts, providing strong support for legal debates and case analysis. We use CAIL2022's Argument Mining dataset to construct our dataset, transforming the identification of focal points of disputes into a multiple-choice question format.
2-9	Event Detection: Given a sentence from a legal judgement document, detect which events are mentioned in this sentence.	Events are the essence of facts in legal cases. Therefore, Legal Event Detection is fundamentally important and naturally beneficial to case understanding and other Legal AI tasks. We construct the test set from the LEVEN dataset(Yao et al., 2022) by sampling sentences corresponding to the top 20 most frequent event types. Multiple events can be mentioned in every sentence.
2-10	Trigger Word Extraction: Given a sentence from a legal judgment document and its corresponding events, predict which words in the sentence triggered these events.	Trigger words directly cause events and are an important feature that determines the event category, providing post-hoc explanation for the event types we identify. Directly identifying trigger words is very difficult, so we simplified this task by providing the events contained in the text along with the text information, examining the ability of LLMs to recognize trigger words related to events. When constructing the trigger word test set, we removed trigger words that were the same as the event type, as well as events with multiple or duplicate trigger words from the LEVEN dataset(Yao et al., 2022), to include as different trigger words as possible.

Table 3: Legal knowledge understanding tasks: definition and data collection details.

nese NLP tasks; and (3) six legal-specific LLMs, which are further fine-tuned on Chinese corpora in the legal domain to improve LLMs' understanding of Chinese laws. Details of these models are shown in Table 5. These 21 models are chosen in part because of their popularity and superior performance,

ID	Definition	Data Collection Details
3-1	Fact-based Article Prediction: Given a fact statement from the legal judgement document, predict which article items should be applied.	When judges make decisions, they usually associate relevant articles with the facts of the case (Ge et al., 2021; Louis et al., 2023). Article prediction can assist judges in quickly locating legal articles related to legal texts. Legal articles are written expressions of legal norms, which are rules and regulations with clear meanings and legal effects.The model needs to deduce potentially applicable legal provisions based on the given case description and related background information. We sample 500 cases from the CAIL2018 dataset for this task.
3-2	Scene-based Article Prediction: Given a described scenario and a related question, predict the corresponding article item.	The CAIL2018 dataset only covers criminal law-related legal provisions. In order to comprehensively evaluate the ability of LLMs to analyze case facts and infer relevant legal provisions, we collected high-quality legal scenario-based question-and-answer data from public sources on GitHub(Liu et al., 2023a). This dataset was generated by inputting legal provisions into chatGPT to construct corresponding scenario-based questions and answers. We manually selected 5,000 question-and-answer pairs with accurate answers from the generated dataset. Based on this, we selected 252 core legal provisions' scenario-based question-and-answer content as the test dataset.
3-3	Charge Prediction: Given fact statement from the legal judgement document and the applied article number, predict the cause of action (charge).	Cause of action is a summary of the nature of the legal relationship involved in a litigation case, as formulated by the people's court. Accurately predicting the cause of action can help improve judicial efficiency and fairness. In the process of filing and hearing cases, accurate prediction of the cause of action can help the court to allocate cases, allocate resources, and arrange trials, thereby improving judicial efficiency and fairness. We sampled 500 pieces of data from the CAIL2018 cause of action prediction dataset for this task.
3-4	Prison Term Prediction w.o. Article: Given fact statement from the legal judgement document, the applied article number and charge, predict the prison term.	Prison term prediction refers to the process of predicting and estimating the possible sentence that a defendant may face during the criminal justice process based on the facts of the case, legal provisions, and relevant guiding precedents. It aims to make reasonable inferences about the length and form of the sentence by comprehensively considering various factors such as the nature of the crime, the circumstances of the offense, the social impact, and the defendant's personal situation. We used the prison term prediction dataset from CAIL2018, removed cases with the death penalty and life imprisonment, and randomly sampled 500 cases as the test dataset. During the process of judges' sentencing, more information is usually taken into account to determine the prison term outcome. We simulated the judge's analysis process by providing the relevant legal provisions and the charge of the case.
3-5	Prison Term Prediction w. Article: Given fact statement from the legal judgement document, the applied article content and charge, predict the prison term.	Large language models typically use retrieval-argument methods to introduce new information. Some publicly available models also include retrieval modules that provide detailed reference information for the model by retrieving legal provisions. We simulated this process, and unlike the previous task where only the legal provision number was provided, we provided the specific content of the legal provision in this task. When constructing the sentence prediction dataset, we appended the content of the legal provisions to the end of the question, allowing the model to complete the sentence prediction task in this scenario.
3-6	Case Analysis: Given a case and a corresponding question, select the correct answer from 4 candidates.	We use the case analysis part from JEC_QA dataset (Zhong et al., 2020) for this task. The case analysis part tests the ability of models to analyze real cases. Models must possess five types of reasoning in order to perform this analysis including word matching, concept understanding, numerical analysis, multi-paragraph reading, and multi-hop reasoning. In order to reduce the difficulty of the test and facilitate the acquisition of answers, we sampled 500 multiple-choice questions from the JEC_QA Case-Analysis part as the testing dataset.
3-7	Criminal Damages Calculation: Given a fact description about a criminal process, predict the amount of money involved in this case.	There are some numerical computing tasks in the process of judicial trials, such as the calculation of the total amount of legal crimes. The total amount of the crime is an important sentencing factor. In some charges such as theft, financial fraud, and bribery, China's laws determine the severity of the sentence based on the amount involved in the case. This task mainly tests the computing ability of LLMs. First, we examine whether the model understands the rules of case amount calculation, and second, we examine whether the model can accurately complete numerical calculations. We selected the LAIC2021 numerical computing task to construct our dataset.
3-8	Consultation: Given a user consultation, generate a suitable answer.	Legal consultation is a way for the public to access legal services, helping people understand legal disputes and seek targeted advice and solutions from professional lawyers as well as receive support and guidance. Some law firms and legal consulting companies also provide online legal consultation services, making it more convenient for people to obtain legal help. We collected legal consultation contents from the Hualv website (see Appendix F on hualv.com). Our dataset contains both the answers to legal consultations and the corresponding legal basis, i.e., legal articles.

Table 4: Legal knowledge application tasks: definition and data collection details.

and in part because they are easily accessible.[10]

4.2 Results and Discussion

Figure 1 shows the overall zero-shot and one-shot performance of each model, where the models are ranked by the scores obtained by macro-averaging their scores over the 20 tasks. As can be seen, Qwen-1.5-72B-Chat and GPT-4 are the best performers. Given the same model size (7B-13B), the Chinese oriented LLMs outperform the multilingual models such as StableBeluga2 and Llama by a significant margin, suggesting the usefulness of pre-training and fine-tuning on Chinese data. Interestingly, the legal-specific LLMs do not necessarily

[10] For some models, we use multiple variants that differ in model size.

Model	Parameters	SFT	RLHF	Access	Base Model	Release Date
Multilingual LLMs						
LLaMA-2-Chat (Touvron et al., 2023)	7/13/70B	✓	✓	Weights	LLaMA-2-7/13/70B	2023/7
StableBeluga2 (Mahan et al., 2023)	70B	✓	✗	Weights	LLaMA-2-70B	2023/7
ChatGPT (OpenAI, 2022)	N/A	✓	✓	API	-	2022/11
GPT-4 (OpenAI, 2023)	N/A	✓	✓	API	-	2023/3
Chinese-oriented LLMs						
ChatGLM2 (Du et al., 2022)	6B	✓	✓	Weights	ChatGLM	2023/6
Ziya-LLaMA (Zhang et al., 2022a)	13B	✓	✓	Weights	LLaMA-13B	2023/5
Baichuan2-Chat (Yang et al., 2023)	7/13B	✓	✗	Weights	Baichuan2-7/13B	2023/6
InternLM2-Chat (Team, 2023)	7/20B	✓	✓	Weights	InternLM2-7/20B	2024/2
Qwen-1.5-Chat (Bai et al., 2023)	7/14/72B	✓	✓	Weights	Qwen-1.5-7/14/72B	2024/4
Legal Specific LLMs						
LexiLaw (Lex, 2023)	6B	✓	✗	Weights	ChatGLM-6B	2023/6
Wisdom-Interrogatory (Wis, 2023)	7B	✓	✗	Weights	Baichuan-7B	2023/8
Fuzi-Mingcha (Wu et al., 2023)	6B	✓	✗	Weights	ChatGLM-6B	2023/9
Lawyer-LLaMA (Huang et al., 2023)	13B	✓	✗	Weights	LLaMA	2023/6
ChatLaw (Cui et al., 2023)	13/33B	✓	✗	Weights	Ziya-LLaMA-13B/Anima-33B	2023/6

Table 5: The LLMs included in our evaluation. For each LLM, the table shows (1) whether the model was trained using supervised fine-tuning (**SFT**) and reinforcement learning from human feedback (**RLHF**); (2) how the model can be accessed (i.e., whether the model parameters are available for download (Weights) or whether the model can only be accessed via an API (API)); (3) the base model; and (4) its release date.

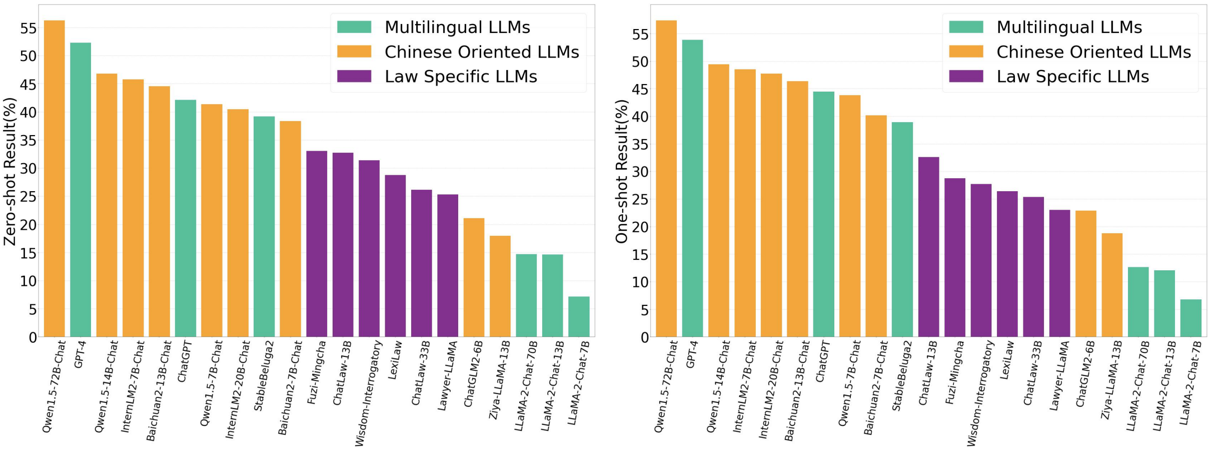

Figure 1: Average performance of the 21 LLMs evaluated on LawBench in the zero-shot (left) and one-shot (right) settings.

outperform the general-purpose Chinese oriented LLMs. A close inspection of the results reveals that existing legal-specific LLMs are based on rather weak foundation models, implying that improved versions of these models may be obtained by fine-tuning a stronger foundation model.

In order to perform a deeper analysis, we show in Table 6 the results of the LLMs on the 20 tasks.[11] For *memorization* tasks, we found that models specifically pre-trained on legal datasets performed the best. For instance, Wisdom-Interrogatory was pre-trained with a large number of laws and regulations, enabling it to accurately repeat legal provisions. Following closely is Qwen-1.5-72B, which, despite being a general-purpose Chinese LLM, likely includes some law-related data within its extensive Chinese training materials as it scores higher than some specialized legal models in Task 1-1. In addition, we observed that GPT-4 could not precisely reproduce statute content: from its responses, it appeared to either believe certain laws did not exist or fabricate statutory content. Despite this lack of precision regarding detailed legislative contents, GPT-4's vast training dataset may have contained much material related to Chinese law, giving it an understanding of Chinese legal concepts.[12] Consequently, GPT-4 ranked second in Task 1-2. In Task 2-2, larger-scale models trained

[11] Owing to space limitations, we only show in this table the results of eight LLMs, which are chosen because they achieved the best result on at least one of the 20 tasks. The per-task results on all models can be found in Appendix K.

[12] The fact that we do *not* know what data the general-purpose LLMs were pre-trained on makes it difficult to analyze their behavior, thus the speculations in our analysis.

Model	Qwen	GPT4	Intern	Intern	Stable	Fuzi	Wisdom	Lexi
Size	72B	N/A	7B	20B	70B	7B	7B	6B
1-1 (Article Recitation)	29.13	15.38	13.03	11.95	14.58	25.22	**43.05**	16.96
1-2 (Knowledge Question Answering)	**76.40**	55.20	50.20	42.00	34.60	7.80	15.40	21.00
All Memorization Tasks	**52.77**	35.29	31.61	26.98	24.59	16.51	29.23	18.98
2-1 (Document Proofreading)	35.01	12.53	**36.78**	33.19	7.70	4.93	30.97	6.24
2-2 (Dispute Focus Identification)	**48.60**	41.65	39.20	43.20	25.57	19.59	7.84	3.30
2-3 (Marital Disputes Identification)	62.05	**69.79**	54.52	54.95	44.20	28.46	36.72	15.60
2-4 (Issue Topic Identification)	39.00	44.00	43.80	**44.20**	39.00	18.60	21.00	22.80
2-5 (Reading Comprehension)	66.47	56.50	47.21	33.45	52.03	**97.59**	35.56	45.39
2-6 (Named-Entity Recognition)	75.53	**76.60**	51.50	12.51	65.54	44.07	57.06	48.74
2-7 (Opinion Summarization)	34.81	37.92	33.60	34.65	39.07	**54.32**	33.34	33.12
2-8 (Argument Mining)	54.40	**61.20**	43.20	11.00	45.80	8.80	10.60	21.60
2-9 (Event Detection)	70.55	**78.82**	63.89	62.83	65.27	16.90	15.98	15.30
2-10 (Trigger Word Extraction)	43.29	**65.09**	36.32	31.44	41.64	7.78	6.24	11.17
All Understanding Tasks	52.97	**54.41**	45.00	36.14	42.58	30.10	25.53	22.32
3-1 (Fact-based Article Prediction)	**72.42**	52.47	63.79	69.98	16.41	25.19	32.84	13.15
3-2 (Scene-based Article Prediction)	29.67	27.54	14.12	12.96	24.52	22.18	32.01	**35.78**
3-3 (Charge Prediction)	**57.07**	41.99	48.91	48.00	22.82	55.93	35.09	39.99
3-4 (Prison Term Prediction w/o Article)	81.32	**82.62**	81.42	81.81	76.06	77.23	80.36	78.08
3-5 (Prison Term Prediction w/ Article)	79.95	81.91	80.11	**83.19**	65.35	75.52	81.10	74.92
3-6 (Case Analysis)	**70.40**	48.60	39.60	17.40	34.40	7.00	15.40	20.80
3-7 (Criminal Damages Calculation)	74.80	**77.60**	55.40	63.20	56.60	47.20	17.40	35.80
3-8 (consultation)	**24.30**	19.65	19.32	17.45	13.39	16.64	20.17	15.82
All Application Tasks	**61.24**	54.05	50.33	49.25	38.69	40.86	39.29	39.29
Overall	**56.26**	52.35	45.80	40.47	39.23	33.05	31.41	28.78
Abstention	2.61	1.50	2.01	2.18	5.00	7.70	4.40	7.90

%abstention 0% 25% 50% 75% 100%

(a) Zero-shot results.

Model	Qwen	GPT4	Intern	Intern	Stable	Fuzi	Wisdom	Laxi
Size	72B	N/A	7B	20B	70B	7B	7B	6B
1-1 (Article Recitation)	25.71	17.21	17.04	21.07	15.03	20.21	**37.41**	15.47
1-2 (Knowledge Question Answering)	**74.40**	54.80	47.00	50.60	36.00	12.80	15.20	14.40
All Memorization Tasks	**50.06**	36.00	32.02	35.84	25.51	16.50	26.30	14.94
2-1 (Document Proofreading)	35.01	18.31	36.78	**46.81**	8.93	2.86	22.16	4.18
2-2 (Dispute Focus Identification)	44.20	**46.00**	40.00	43.40	15.00	2.40	10.00	15.40
2-3 (Marital Disputes Identification)	65.35	**69.99**	49.55	54.65	41.76	17.44	24.29	21.49
2-4 (Issue Topic Identification)	40.60	**44.40**	41.80	43.20	38.00	8.80	13.40	27.20
2-5 (Reading Comprehension)	78.46	64.80	61.62	70.72	53.55	**93.35**	13.23	41.64
2-6 (Named-Entity Recognition)	73.83	**79.96**	64.95	27.05	64.99	42.28	40.10	31.54
2-7 (Opinion Summarization)	42.11	40.52	37.12	37.90	**45.06**	31.43	39.71	34.57
2-8 (Argument Mining)	57.60	**59.00**	44.80	22.60	37.60	11.40	0.20	6.00
2-9 (Event Detection)	74.71	**76.55**	66.54	67.69	65.89	21.26	15.81	19.84
2-10 (Trigger Word Extraction)	37.35	**65.26**	40.18	42.85	40.54	7.04	4.02	8.36
All Understanding Tasks	54.92	**56.48**	48.33	45.69	41.13	23.83	18.29	21.02
3-1 (Fact-based Article Prediction)	73.79	53.20	64.15	**73.85**	16.87	3.86	20.02	15.41
3-2 (Scene-Based Article Prediction)	**36.10**	33.15	29.35	20.51	32.44	32.96	23.33	33.94
3-3 (Charge Prediction)	**60.01**	41.30	51.03	51.00	23.07	43.60	39.22	34.03
3-4 (Prison Term Prediction w/o Article)	80.77	**83.21**	80.11	81.57	75.80	78.95	81.16	73.66
3-5 (Prison Term Prediction w/ Article)	79.11	**82.74**	80.21	82.04	63.59	79.00	81.57	70.93
3-6 (Case Analysis)	**68.80**	49.60	41.40	32.00	33.00	13.80	13.00	12.20
3-7 (Criminal Damages Calculation)	75.00	**77.00**	56.60	62.60	56.00	38.20	40.40	33.20
3-8 (Consultation)	**24.67**	19.90	20.42	22.79	16.24	13.95	20.67	14.68
All Application Tasks	**62.28**	55.01	52.91	53.29	39.63	38.04	39.92	36.01
Overall	**57.38**	53.85	48.53	47.74	38.97	28.78	27.74	26.41
Abstention	2.45	1.60	2.25	2.06	5.70	8.40	1.80	8.80

%abstention 0% 25% 50% 75% 100%

(b) One-shot results.

Table 6: Per-task zero-shot and one-shot results of the LLMs. The strongest results are **boldfaced**.

using more data tended to retain more legal knowledge. In particular, the enhanced Qwen-1.5-72B model and GPT-4 are the best performers.

For *understanding* tasks, GPT-4 and Qwen-1.5-

72B achieve comparable performance, with the former offering slightly stronger results on average. However, there are some tasks on which one of them performs substantially better than the other. First, Qwen-1.5-72B substantially outperforms GPT-4 on two tasks, 2-1 and 2-5. In Task 2-1 (legal text correction), while GPT-4 demonstrates strong capabilities in correcting incorrect parts of a legal text, it also unnecessarily alters some correct sections, resulting in a lower score for this task. In contrast, Qwen-1.5-72B is better at Chinese legal terminology, effectively rectifying necessary modifications without overcorrecting unnecessary areas. Second, in Task 2-5 (reading comprehension), Qwen-1.5-72B performs much better than GPT-4. While both models are reasonably good at answering questions, Qwen-1.5-72B's responses are more concise, granting it a higher score than GPT-4. Surprisingly, Fuzi achieves unusually high performance on this task, possibly because it was fine-tuned on related tasks. Given its near-perfect performance, it is even possible that it was fine-tuned on the test data used for this task. Finally, for Task 2-10 (case-related information trigger word detection), GPT-4 substantially outperforms Qwen-1.5-72B. While Qwen-1.5-72B can identify related triggers, it struggles with precise identification, often outputting irrelevant triggers. In contrast, GPT-4 exhibits stronger comprehension abilities and refrains from producing excessive content.

For *application* tasks, Qwen-1.5-72B significantly outperforms GPT-4 in the zero-shot setting on Tasks 3-1, 3-3, and 3-6, making it the best-performing model. For Task 3-1 (fact-based article prediction), our analysis shows that while GPT-4 can accurately identify legal provisions for common cases such as robbery and theft crimes, it fails to effectively recognize relevant provisions for less common offenses like illegal logging. This leads to a lower accuracy rate compared to Qwen-1.5-72B.

For Task 3-3 (charge prediction), we also observed that GPT-4 precisely identifies offenses like theft and robbery but struggles with others such as counterfeiting financial documents or insurance fraud. For instance, when faced with defamation by false accusation, GPT-4 might generate inaccurate charges such as framing someone.

For Task 3-6 (legal case analysis), Qwen-1.5-72B excels at synthesizing legal knowledge and concepts, achieving the best results. GPT-4 similarly leverages its strong analytical capabilities and understanding of legal concepts, enabling it to perform the second best in this domain.

To gain further insights into these models, we measure their *abstention rate* on each task, which is defined as the percentage of queries (i.e., test instances) they failed to answer. We found that some general-purpose LLMs, such as Qwen-1.5-72B and GPT4, have better instruction-following capabilities, which makes their abstention rates low in most tasks (see Table 6). In contrast, fine-tuned models often suffer from a lack of diversity in the training data they are fine-tuned on, leading to a degradation of their ability to follow instructions. For example, legal-specific models like Fuzi, Wisdom, and LexiLaw frequently produce irrelevant content because they do not fully comprehend our test tasks, resulting in high abstention rates.

In sum, these results demonstrated that while we are still a long way from obtaining reliable results from LLMs in Chinese legal tasks, they revealed that Qwen-1.5-72B and GPT-4 are the best performers on average. Equally importantly, our analysis provided reasons why one model performs substantially than the other on these tasks, suggesting areas of improvements for these models in future work.[13]

5 Conclusion

We presented *LawBench*, the first benchmark composed of 20 tasks that is aimed to assess LLMs in performing legal-related tasks under the Chinese civil law system. We provided a structured taxonomy of the skill set required for legal-related tasks through which the 20 tasks were divided into three groups that correspond to three cognitive dimensions in Bloom's Taxonomy, namely memorization, understanding and application. We assessed the performance of 21 LLMs on the 20 tasks. Our results demonstrated that current LLMs were still unable to give meaningful judicial aid, and their scores on most tasks were often poor. Nevertheless, we conducted the first *comparative analysis* of the results in an attempt to reveal the strengths and weaknesses of these LLMs on the 20 legal tasks. Our analysis revealed that while fine-tuning open-source LLMs on legal data resulted in some improvements, they still offered inferior performance to GPT-4. As the legal field is highly professional, much of the data used in practical applications is confidential. Therefore, we believe that the release of LawBench and our analysis can serve as a solid foundation for future research in this area.

[13] A detailed analysis can be found in Appendix K.

Limitations

The majority of our datasets are acquired through sampling publicly available data on the . Although we have made efforts to select the newest versions of these datasets, there can still be risks of test data leakage given that existing LLMs have been exhaustively trained on massive amounts of Internet data. It is possible that the LLMs that have been explicitly trained on these task formats, or even the exact test data, can achieve exceptionally high scores (Schaeffer, 2023). We will seek more principled ways to prevent data contamination in the future.

Another limitation concerns the answer extraction methods and the evaluation metrics we employed for generative tasks. Although we have hand-engineered task-specific rules to extract the answers, there can still be cases where a rule fails to match. For generative tasks, we only use Rouge-L to evaluate the model predictions, which is not ideal in the sense that the Rouge-L scores may not correlate well with human judgments of answer quality. Currently, there is a lack of automated methods for effectively evaluating model predictions from legal tasks. We plan to consider training an evaluation model tailored for legal tasks in the future, or experiment with LLM-based evaluations (Liu et al., 2023b; Yu et al., 2024).

References

2023. Lexilaw. https://github.com/CSHaitao/LexiLaw.

2023. Wisdominterrogatory. https://github.com/zhihaiLLM/wisdomInterrogatory.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *Preprint*, arXiv:2307.16877.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *Preprint arXiv:2309.16609*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

OpenCompass Contributors. 2023. OpenCompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *Preprint*, arXiv:2306.16092.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *Preprint*, arXiv:2204.04859.

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). *Preprint arXiv:2310.05620*.

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, Dublin, Ireland. Association for Computational Linguistics.

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

Neel Guha, Julian Nyarko, Daniel 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. New Orleans, Louisiana.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer LLaMA technical report. *Preprint arXiv:2305.15062*.

Yunyun Huang, Xiaoyu Shen, Chuanyi Li, Jidong Ge, and Bin Luo. 2021. Dependency learning for legal judgment prediction with a unified text-to-text transformer. *Preprint arXiv:2112.06370*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*, pages 32537–32551, New Orleans, Louisiana.

David R. Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into Practice*, 41(4):212–218.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023a. LawGPT: Chinese legal dialogue large language model. https://github.com/LiuHC0428/LAW_GPT.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. MMBench: Is your multi-modal model an all-around player? *Preprint arXiv:2307.06281*.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023. Finding the law: Enhancing statutory article retrieval via graph neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2761–2776, Dubrovnik, Croatia. Association for Computational Linguistics.

Dakota Mahan, Ryan Carlow, Louis Castricato, Nathan Cooper, and Christian Laforte. 2023. Stable beluga models. https://huggingface.co/stabilityai/StableBeluga2.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E Ho, and Joel Niklaus. 2023. Scale: Scaling up the complexity for advanced language model evaluation. *arXiv preprint arXiv:2306.09237*.

Rylan Schaeffer. 2023. Pretraining on the test set is all you need. *Preprint arXiv:2309.08632*.

InternLM Team. 2023. InternLM: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM-techreport.

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. *Preprint arXiv:2307.09288*.

Arianna Trozze, Toby Davies, and Bennett Kleinberg. 2024. Large language models in cryptocurrency securities cases: can a gpt model meaningfully assist lawyers? *Artificial Intelligence and Law*, pages 1–47.

Shiguang Wu, Zhongkun Liu Liu, Zhen Zhang 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.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *Preprint arXiv:2309.10305*.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale Chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 183–201, Dublin, Ireland. Association for Computational Linguistics.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu GONG, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Xu Bin, Jie Tang, and Juanzi Li. 2024. KoLA: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

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

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

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022b. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States. Association for Computational Linguistics.

Henry R Zheng. 1986. China's new civil law. *The American Journal of Comparative Law*, 34(4):669–704.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9701–9708, New York, New York. AAAI Press.

A Tasks in LAiW

Using the task ids mentioned in Table 1, the six tasks that are present in both LawBench and LAiW are 2-2, 3-1, 2-6, 2-7, 3-3, 1-2, and 3-6 (note that they merge 1-2 and 3-6 into one task, Legal QA). Their Prison Term Prediction task is similar to our Tasks 3-4 and 3-5; their Case Understanding task is similar to Task 2-5, and their Legal Consultation task is similar to Task 3-8. The remaining tasks, Element Recognition, Case Recognition, Similar Case Matching, Civil Trial Prediction, and Judicial Reasoning Generation, are not present in LAiW. We refer the reader to the LAiW paper for details.

B Legal Jurisdictions

In this section, we provide a discussion of the differences between two law jurisdictions, common law jurisdictions and civil law jurisdictions.

Generally speaking, in the civil law system, judges consider the provisions of the written law and conduct trials according to the provisions of written law. In contrast, in the common law system judges consider the precedent of similar cases in the past, and compare this case with the precedent to find the principle and basis of this case.

Given the aforementioned differences, when designing the Legal NLP evaluations for a civil law system, there should be an emphasis on assessing a model's understanding of abstract legal articles. This includes the ability to associate relevant legal articles with given cases and to analyze the outcomes of cases based on these articles. In LawBench, Task 3-2 Scene-based Article Prediction and 3-1 Fact-based Article Prediction are designed to assess a model's ability to analyze relevant articles and Task 3-5 Prison Term Prediction w. Article examines a model's ability to reason with the results based on the articles and the given case. In contrast, when designing the Legal NLP evaluations for a common law system, it is essential to scrutinize a model's ability to recognize similar cases, as well as its understanding of the reasoning behind judicial decisions and the material facts involved. For example, the overruling task in LegalBench evaluates an LLM's ability to identify when a sentence from a judicial opinion overrules a previous case.

C Motivation for using Bloom's Taxonomy

The Bloom's Taxonomy framework provides a valuable measure for the process of learning knowledge. Some studies utilize this framework to assess large language models' grasp of world knowledge[1]. Since the tasks related to law are knowledge-intensive, we think this framework can effectively gauge the extent to which large language models have mastered legal knowledge. Other evaluation frameworks categorize tasks based on their level of difficulty, merely analyzing the capabilities of large language models in legal tasks without assessing the extent to which the models grasp legal knowledge. Using Bloom's Taxonomy can not only define the scope of desired abilities but also model the inherent connections between the evaluated abilities, which allows for diagnostic insights on how to acquire and improve these abilities. For instance, in the LawBench 1-1 task (Article Recitation), we examine the model's grasp of laws and regulations, while in the LawBench 3-2 task (Scene-based Article Prediction), the model needs to have a thorough understanding of the legal provisions in order to accurately recall and recite the relevant articles. If the model has a poor memory of legal knowledge, it would likewise perform poorly on the 3-1 task. Within Bloom's Taxonomy, the memorization of legal knowledge is considered a fundamental skill. We believe that large language models, equipped with extensive legal knowledge, can perform legal tasks more effectively.

Considering the difficulty in demarcating boundaries between higher-order abilities within Bloom's Taxonomy, it proves challenging to categorize a task into a higher level. That is the reason why we simplify and select three widely accepted cognitive processes in Bloom's learning theory for organizing the tasks in LawBench.

D Details of Task Instruction

D.1 Legal Knowledge Memorization Tasks

Table 7 presents examples of Task 1-1, while Table 8 illustrates instances relevant to Task 1-2.

D.2 Legal Knowledge Understanding Tasks

Tables 9 through Table 18 provide examples for the Legal Knowledge Understanding Tasks.

D.3 Legal Knowledge Application Tasks

Tables 19 through Table 26 provide examples for the Legal Knowledge Application Tasks.

E Prompt Engineering

In this section, we discuss our process of prompt engineering and the effort spent on each model and task.

When designing these prompts, we were inspired by the Prompt Engineering procedure in LegalBench. We crafted the instructions for each task manually and chose zero to one example as an illustration. When crafting these prompts, we adhere to principles of clarity and conciseness. We endeavor to construct prompts that clearly articulate the question intent while being succinct, thus lowering the difficulty for the model to comprehend the problem. Furthermore, for each question, we specify an output format to facilitate the extraction of answers. To ascertain whether the instructions we've devised are broadly applicable to most models, we input the crafted prompts into the models we are testing and check whether each model can follow the instructions to produce answers that conform to the specified format. We adjust our prompts based on the outputs from the majority of the models. During testing, we concatenate the issues with the corresponding chat templates for each model. When choosing the wording for our prompts, we tested between using specialized vocabulary and plain language descriptions of tasks. We reached the same conclusion as LegalBench: we found that plain language descriptions of tasks enabled large language models to complete tasks more effectively compared to using professional jargon. Inspired by the "Reliance on latent knowledge" in LegalBench, we also evaluated whether providing specific interpretations of legal statutes could enhance the model's performance, as seen in LawBench's Task 3-5 "Prison Term Prediction with Article," assessing the model's ability to complete tasks when given the specific content of statutes. Additionally, we investigated the influence of different examples on the final outcomes. We found that the model's results varied with the examples provided. We concur with the perspective in LegalBench that these models are sensitive to the prompts and the instructions given. We randomly selected an example as our final test case. Despite our significant efforts to improve the model's performance as much as possible, there is still substantial room for enhancement. Our findings serve as a lower-bound on performance.

F Details of Data Source

- FLK: FLK is a national database [14] comprehensively collects Chinese laws and regulations, including the Constitution of the People's Republic of China, civil law, local regulations, etc.

- JEC_QA: JEC_QA(Zhong et al., 2020) is the largest question answering dataset collected from the National Judicial Examination of China. All data can be accessed from http://jecqa.thunlp.org/.

- CAIL: CAIL(Challenge of AI in Law) [15] is a competition website related to law, which aggregates many test tasks in the Chinese judicial field.

- LAIC: LAIC(Legal AI Challenge) [16] is another competition website about legal tasks that offer different competition tasks distinct from those on CAIL.

- AIStudio: AI Studio [17] is a learning platform for deep learning that offers extensive open datasets, including some relevant to legal. We constructed 2-3 tasks from the public

[14] https://flk.npc.gov.cn/

[15] http://cail.cipsc.org.cn/

[16] https://laic.cjbdi.com/

[17] https://aistudio.baidu.com/index

INSTRUCTION: Answer the following questions, just give the content of the law directly:

QUESTION: What is the content of Article 257 of the Criminal Law?

ANSWER: Anyone who uses violence to interfere with the marriage freedom of others shall be sentenced to fixed-term imprisonment of no more than two years or detention. Anyone who commits the crime mentioned in the preceding paragraph and causes the death of the victim shall be sentenced to fixed-term imprisonment of no less than two years and no more than seven years.

Table 7: The instruction and an example of Task 1-1 Article Recitation.

INSTRUCTION: Please apply your legal knowledge to select the correct answer from A, B, C, or D and write it between [Correct Answer] and <eoa>. For example, [Correct Answer] A <eoa>. Please strictly follow this format when answering.

QUESTION: According to the laws of our country, in the process of trial of foreign-related divorce cases accepted by our courts, which of the following should be used as the basis for determining the validity of the marriage? A: The law of the place where the marriage was concluded; B: The law of the parties' home country; C: The law of the parties' place of residence; D: The law of the court.

ANSWER: [Correct Answer] A<eoa>

Table 8: The instruction and an example of Task 1-2 Knowledge Question Answering.

dataset [18] on this platform.

- CrimeKgAssitant: CrimeKgAssistant [19] is an open-source crime assistant github project. This dataset consists of 856 pieces of crime knowledge graphs, a 2.8 million crime prediction training dataset, 200k legal Q&A pairs, and a 13-category topic classification for these 200k legal consultation questions.

- LawGPT: LawGPT(Liu et al., 2023a) is an open-source Chinese legal large model github project. In this project, they public the training dataset but do not release the trained Chinese legal large language model. The training dataset includes scenario dialogues between lawyers and users, some of which are cleaned from the publicly legal data CrimeKgAssistant, while others were generated by utilizing ChatGPT to conceive specific question-answering scenarios based on 9,000 key legal provisions thereby ensuring that the generated dataset has concrete legal grounds.

- hualv.com: hualv.com [20] is an online platform dedicated to providing legal consultation services, where numerous real users and lawyers engage in daily interactions by asking and answering questions. Topics of these conversation data range from marriage-related questions, labor disputes, contract controversies, etc. All data on this platform is public available and can be accessed through web scraping.

- LEVEN: LEVEN(Yao et al., 2022) is the largest Legal Event Detection(LED) dataset with 8, 116 legal documents and 150, 977 human annotated event mentions in 108 event types. Not only charge-related events, LEVEN also covers general events, which are critical for legal case understanding but neglected in existing LED datasets.

G Details of Evaluation

Answer Extraction Most of the tasks require the prediction to be in the standard format in order to compare with the ground truth, we define a set of task-specific rules to extract the answer from the model prediction.

- Article Number Extraction (3-1): this type of tasks requires us to extract the article numbers predicted by the model. To do this, we use the delimiter "、" to separate the prediction text into chunks of text, and then the cn2an[21] library is used to convert the Chinese numerals to Arabic numerals within each of those chunks. Using a regular expression, we extract the converted Arabic numerals as the expected article numbers; if more than one number appears in the same chunk, only the

[18] https://aistudio.baidu.com/datasetdetail/181754

[19] https://github.com/liuhuanyong/CrimeKgAssitant

[20] www.66law.com

[21] https://github.com/Ailln/cn2an

INSTRUCTION: Correct the spelling, redundancy, omission, and disorder errors of the words in the following legal document sentences, and minimize modifications to the original sentences while preserving their semantics. Only the modified sentences need to be provided in the answer, strictly following this format. Sentence:

QUERY: The above-mentioned procedure for collecting the evidence for the appeal is legal, and the content is objective and true, which is sufficient to establish the accusation.

ANSWER: The above-mentioned procedure for collecting evidence is legal, and the content is objective and true, which is sufficient to establish the facts of the accusation.

Table 9: The instruction and an example of Task 2-1 Document Proofread.

INSTRUCTION: Determine the category of the dispute focus contained in the sentence. Each sentence contains only one dispute focus category. Categories include: litigation parties, rent situation, interest, principal dispute, liability determination, liability division, loss determination and handling, whether the original judgment is appropriate, contract effectiveness, property division, liability assumption, admissibility of appraisal conclusions, statute of limitations, breach of contract, contract termination, hit-and-run. Write the answer between [Dispute Focus] and <eoa>, such as [Dispute Focus] Principal Dispute <eoa>.

QUERY: Sentence: The plaintiff alleges: I have a father-son relationship with the defendant. The house located at Unit 321, Building 7, 3 North Xiaojie, Sanlitun North, Chaoyang District, Beijing (hereinafter referred to as the 'involved property') was allocated to me by the Ministry of Foreign Affairs Diplomatic Service Bureau during my time at TIME. My unit underwent housing reform, and I purchased the property at TIME. I have been living in the involved property from TIME to the present. In TIME, the defendant started giving various reasons for not allowing me to reside in the involved property, which raised my suspicions. In TIME, I went to ORG to inquire about the status of the involved property and discovered that in TIME, the defendant deceived me into signing an 'Existing Property Sale Contract' and processed the property transfer procedures at the Housing Management Bureau, transferring the involved property into the defendant's name. I believe I have never sold the involved property, and the defendant deceived me into signing the 'Existing Property Sale Contract.' The defendant also did not actually pay the purchase price specified in the property sale contract. Therefore, I have filed a lawsuit to the court requesting the cancellation of the 'Property Sale Contract' signed between me and the defendant and the transfer of the property to my name. The defendant argues: The plaintiff and I signed the contract voluntarily. The involved property is state-owned property, and the sale of state-owned property requires strict approval procedures. All the documents were signed by the plaintiff himself, and I also fully paid the purchase price. Now the plaintiff is backtracking because several other children are manipulating the elderly, and the elderly are confused. I do not agree with the plaintiff's lawsuit request.

ANSWER: Dispute focus category: contract effectiveness <eoa>

Table 10: The instruction and an example of Task 2-2 Dispute Focus Identification.

first number is extracted. All extracted numbers are combined to form the final set of predictions.

- Prison Term Extraction (3-4, 3-5): for this type of tasks, we need to extract the predicted prison terms from the prediction text. To begin, we use cn2an to convert all the Chinese numerals in the prediction to Arabic numerals; we then extract digits that are followed by time intervals in Chinese, such as “个月” (month) and “年” (year). The extracted prison terms are normalized to months, meaning that the numbers appearing before “年” will be multiplied by 12. Note that the time unit in the ground truth answer is also month.

- Criminal Damages Extraction (3-7): We extract all numbers appearing in the prediction text using regular expression. The set of of the extracted numbers is considered as the predicted criminal damages.

- Named-Entity Recognition (2-6): We find all occurrences of entity types from the model prediction, then obtain the substring from its occurrence to the delimiter “\n”, then apply a regular expression to extract the entity value.

- Trigger Word Extraction (2-10): We split the model prediction by the delimiter “；”, then treat the split array as a list of extracted key words.

- Option Extraction (1-2, 2-2, 2-3, 2-4, 2-8, 2-9, 3-3): this type of task is similar to selecting the correct options from a list of options in a multiple-choice task. We run through all possible options and check if they appear in the prediction text. The set of options that occur in the prediction text is collected and used for evaluation.

- Others (1-1, 2-1, 2-5, 2-7, 3-2, 3-8): we take the model prediction as the answer without performing any extraction step.

Metrics After the answer extraction phase, we compute the final metric based on the extracted

INSTRUCTION: Please categorize these sentences based on tags, which include: having children after marriage, custody of children with limited capacity, having joint property, paying child support, dividing real estate, post-marital separation, filing for divorce for the second time, monthly payment of child support, granting divorce, having joint debt, personal property before marriage, legal divorce, failure to fulfill family obligations, existence of non-marital children, appropriate assistance, failure to comply with divorce agreements, compensation for damages, separation due to emotional discord for over two years, children living with non-custodial parent, personal property after marriage. Please write the answer between [category] and <eoa>, for example, [category] division of real estate, joint property <eoa>. Please strictly follow this format.

QUERY: In 2014, Wang X sued and demanded a share of the down payment of 150,000 yuan on the grounds that he and Xiao X had not divided the down payment of the building when they agreed to divorce.

ANSWER: [category] dividing real estate, having joint property <eoa>

Table 11: The instruction and an example of Task 2-3 Marital Disputes Identification.

INSTRUCTION: Please determine the category for the following consultation. Each consultation belongs to only one category, which includes: Marriage and Family, Labor Disputes, Traffic Accidents, Debt Collection and Debt Disputes, Criminal Defense, Contract Disputes, Real Estate Disputes, Infringement, Corporate Law, Medical Disputes, Demolition and Resettlement, Administrative Litigation, Construction Engineering, Intellectual Property Rights, Comprehensive Consultation, Personal Injury, Foreign-related Law, Maritime Law, Consumer Rights, Mortgage and Guarantee. Please write the answer between [Category] and <eoa>, for example [Category] Marriage and Family <eoa>. Please strictly follow this format when answering.
Consultation:

QUERY: It has been two years since we separated and have not been together. Will the court automatically grant a divorce? The other party has sued for the dowry money, but the case has not been settled. Will this affect the other party's ability to remarry and obtain a marriage certificate?

ANSWER: Marriage and Family

Table 12: The instruction and an example of Task 2-4 Issue Topic Identification.

answer. We defined 7 different metrics in total to measure different types of tasks:

- **Accuracy**: Accuracy is a binary score that performs exact match between the model prediction and the gold answer. This applies to all single-label classification tasks including task 1-2, 2-4, 2-8, 3-6, and the regression task 3-7. For SLC tasks, if multiple valid answers are extracted from the model prediction, then we always treat it as wrong [22].

- **F1**: When there are multiple output labels, F1 score measures the harmonic mean of the precision and recall. This applies to all multi-label classification tasks including task 2-2, 2-3, 2-9, 3-1 and 3-3.

- **rc-F1**: rc-F1 is the F1 score tailored for the reading comprehension task 2-5. It treats every token as a label, removes punctuation, stories, extra whitespace, performs other necessary normalizations then compute the F1 score. We adopt the official script from CAIL2019 to compute the instance-level rc-F1 score [23].

- **soft-F1**: For extraction tasks 2-6 and 2-10, the output is a set of phrases. Instead of using the standard F1 score, we use a soft version by replacing the phrase-level exact match with the rc-F1 score, then computing the F1 on top of it. We find using the soft version helpful since LLMs often use wording choices different from the ground truth.

- **nLog-distance**: For the prison term prediction tasks 3-4 and 3-5, we evaluate them with the normalized log distance (nLog-distance) to capture the continuity of prison terms. We compute the logarithm of the difference between the extracted and gold answer, then normalize it between 0 and 1 for better compatibility with other metrics.

- **F0.5**: For the document proofreading task 2-1, we use the F0.5 metric to evaluate it. The F0. 5 score gives more weight to precision than to recall we want to prevent introducing more false positives than identify every other error in proofreading (Zhang et al., 2022b). We use the ChERRANT toolkit to align the

[22] For the criminal damages calculation task, we treat the model prediction correct as long as one of the extracted answers match the ground truth as we find LLMs often output the whole calculation process.

[23] https://github.com/china-ai-law-challenge/CAIL2019/tree/master

INSTRUCTION: Please answer the questions based on the provided paragraph as concisely as possible.

QUERY: Paragraph: Based on the valid evidence submitted by the plaintiff to the court and the plaintiff's oral statements in court, this court confirms the following facts in the case: The defendant Li2 and the defendant Li x3 are mother and daughter. On September 26, 2014, the defendant Li2 borrowed 50,000 yuan from Sang x0, and issued a promissory note stating: 'I have borrowed 50,000 yuan in cash from Sang x0 today. The loan term is 3 months, and I will repay it on time every month with an interest of 2,500 yuan per month. If I cannot repay it at the due date, I am willing to use all my assets to settle the debt and take full responsibility. Borrower: Si x1. Spouse: Li2. Willing to take full responsibility.' On November 13, 2014, with the guarantee of the defendant Li x3, the defendant Li2 borrowed 150,000 yuan from Sang x0 and issued a promissory note stating: 'I have borrowed 150,000 yuan in cash from Sang x0 today. The loan term is three months, and I promise to pay the principal and interest monthly and return it on time. I hereby promise. Borrower: Li2, Guarantor: Li x3, Spouse: Si x1.' On the same day, Li2 issued a repayment guarantee commitment, and Si x1 signed it as a spouse. At the same time, Li2 and Si x1 issued a mortgage commitment, promising to use the existing house, car, and any other property in Zone 13 of Dongjiang New Village, Dongjiang Town, as collateral for this loan. However, both parties did not complete the mortgage registration procedure. After the loan maturity date, the defendant Si x1, besides paying 22,000 yuan in interest to the plaintiff, did not repay the remaining principal and interest. The plaintiff then sued in this court, requesting: 1. The defendant Li2 and Si x1 to repay the principal of the loan, which is 200,000 yuan; 2. The defendant Li2 and Si x1 to repay the loan interest at a rate of 2% per month; 3. The defendant Li x3 to assume the guarantee liability within the range of 150,000 yuan; 4. The defendant is to bear the litigation costs. Furthermore, it was found that the defendant Li2 and the defendant Si x1 did not complete the marriage registration procedure. Question: Who borrowed money from the lender?

ANSWER: Answer: Si x1, Li2

Table 13: The instruction and an example of Task 2-5 Reading Comprehension.

INSTRUCTION: Based on the provided entity types, extract entity information from the sentence. The entity types include: suspected criminals, victims, stolen currency, item value, theft profit, stolen items, crime tools, time, location, and organizations. List the entity information one by one.

QUERY: Sentence: After the case was solved, the public security organ lawfully returned the seized mobile phones to the victims, Yan and Xiao.

ANSWER: Victims: Yan, Xiao; stolen items: the seized mobile phones; organizations: the public security organ

Table 14: The instruction and an example of Task 2-6 Name Entity Recognition.

extracted and gold answer before computing the F0.5 score [24]. As the alignment can take too long to respond for very bad generations, we add a time-out of 10 seconds. If a time-out happened, then the prediction is assigned a score of 0.

- **Rouge-L**: For other generation tasks 1-1, 2-7, 3-3 and 3-8, we use the Rouge-L score to evaluate them. Rouge-L is a commonly used metric in generation tasks. It takes into account sentence-level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically to compare the extracted and gold answers (Lin, 2004).

Several large language models may decline to respond to legal-related inquiries due to security policies or simply fail to follow the instructions. To capture this issue, we also report the **abstention rate** of LLMs in each task (how often an LLM abstains to answer). An abstention happens if an answer cannot be extracted from the model prediction. The abstention rate does not apply to task 2-5 and all generation tasks since they do not need the answer extraction step.

H Details of Models

Multilingual LLMs We consider 6 open-source multilingual models: LLaMA-2-Chat-7B / 13B / 70B, StableBeluga2. In addition, two commercial models, ChatGPT and GPT-4, are included.

Chinese Oriented LLMs A number of Chinese-oriented LLMs are proposed to enhance Chinese comprehension. Their typically perform better than multilingual models on Chinese NLP tasks. We include 3 open-sourced, Chinese-oriented LLMs families in our evaluation. We also include two widely used LLMs which are ChatGLM2 and Ziya-LLaMA:

- InternLM2 families (Cai et al., 2024): InternLM2 series contains two model sizes: 7B and 20B. These models are pretrained on massive general domain textual data, programming language-related data, and long textual data. Subsequently, they are aligned to elicit

[24] https://github.com/HillZhang1999/MuCGEC/tree/main/scorers/ChERRANT

INSTRUCTION: Here is a news report, please provide a summary of this report in one sentence.

QUERY: According to a report by Xinmin Evening News (Reporter Xia Yun), at about 9 am today, some citizens reported to the journalist that a unit on Cao'an Road in Jiading District, Shanghai had a sudden fire, with thick smoke at the scene; the fire department arrived promptly and the fire was still being put out by the time of the report, with no casualties reported; the source is the official Weibo account of Shanghai Fire Rescue Bureau.According to witnesses, when the fire broke out, there was a large amount of smoke above the unit, and some drivers on the Jiamin elevated road also saw the smoke; after receiving the alarm, the police, fire and other departments rushed to the scene for disposal; as of 9:30 am, the disposal work was still in progress; according to the official Weibo of Shanghai Fire Rescue Bureau, at about 8:33 am on March 24, a company in Cao'an Road, Jiading District had a fire, and the fire department arrived at the scene for disposal, with the fire now under control and no casualties reported.

ANSWER: At around 9 a.m. this morning, a unit on Cao'an Road in Jiading District, Shanghai had a sudden fire, with thick smoke at the scene.

Table 15: The instruction and an example of Task 2-7 Opinion Summarization.

INSTRUCTION: Based on a given prosecution point of view, please select a point of view that can form a controversial point of view from the five defense candidate points of view A, B, C, D, and E, and write the answer in [Correct Answer] and <eoa>. For example [correct answer]A<eoa>. Please answer strictly according to this format.

QUERY: Sentence: The private prosecutor Yu XX alleged that at around 17:00, the defendants Bao XX, Han XX, and Han XX came to my house and beat me on the head with a wooden stick on the grounds that I had detained their cattle. After I was injured, I was hospitalized at the Xing'an League Mongolian Medical Hospital for 73 days. A: The defender provided the court with a copy of Lv's testimony, requesting the court to declare the defendant Lv innocent and reject the victim's incidental civil claim. B: I think I am not guilty. C: The defendant Bao Moumou claimed that I injured the private prosecutor with a whip and agreed to compensate for the medical expenses. D: 2. The victim's incidental civil lawsuit should be dismissed. E: In addition, disability compensation should not be included in the scope of judgment compensation.

ANSWER: [Correct answer]C<eoa>

Table 16: The instruction and an example of Task 2-8 Argument Mining.

their capabilities and guide LLMs to serve as helpful and harmless AI assistants.

- Baichuan2 families (Yang et al., 2023): Baichuan2 series contains two model sizes: 7B and 13B. These models trained from scratch, on 2.6 trillion tokens.

- Qwen-1.5 families (Bai et al., 2023): Qwen-1.5 series contains three model sizes: 7B, 14B, and 72B. These models have undergone extensive training using up to 3 trillion tokens of diverse texts and codes, encompassing a wide range of areas. These model have consistently demonstrated superior performance across a multitude of downstream tasks, even when compared to their more significantly larger counterparts.

- ChatGLM2 (Du et al., 2022): ChatGLM-6B is an open bilingual language model based on General Language Model (GLM) framework(Du et al., 2022), with 6.2 billion parameters. ChatGLM-6B uses technology similar to ChatGPT, optimized for Chinese QA and dialogue. The model is trained for about 1T tokens of Chinese and English corpus, supplemented by supervised fine-tuning, feedback bootstrap, and reinforcement learning with human feedback. With only about 6.2 billion parameters, the model is able to generate answers that are in line with human preference.

- Ziya-LLaMA (Zhang et al., 2022a): The Ziya-LLaMA-13B-v1 is a large-scale pre-trained model based on LLaMA with 13 billion parameters. They continue pertaining on 125B tokens data. It embraces a strategy of curriculum learning and incremental training during the supervised fine-tuning. To further improve the overall performance of the model, enabling it to fully understand human intentions, reduce "hallucinations" and unsafe outputs, they conducted Human-Feedback Training (HFT) based on the model fine-tuned with instructions.

Legal Specific LLMs Certain Chinese-oriented LLMs are further fine-tuned on Chinese corpus in legal domain to improve LLMs' understanding of Chinese laws. They are of particular interest to us; through our benchmark, we can rigorously gauge their true advance compared to general-purpose LLMs and identify their limitations. Here, we provide detailed descriptions of these models:

INSTRUCTION: Here is a sentence from a legal text. Please determine which events are triggered by this sentence. The events include: Payment/Disbursement, Deception, Search/Seizure, Demand/Request, Sale, Purchase, Profit, Arrest, Appraisal, Consent/Acceptance, Confession, Contact, Aid/Rescue, Rent/Borrow, Injury, Forgery, Prostitution, Bodily Harm, Compensation, Repayment/Refund. Please list the triggered events directly, separated by semicolons, for example, "Payment/Disbursement; Sale". Please answer strictly in this format. Sentence:

QUERY: Xu Moujia (already sentenced) was the Party Branch Secretary of Shanqian Village at the time, responsible for assisting the Changjie Town People's Government in handling policy matters and construction work related to the project.

ANSWER: Aid/Rescue

Table 17: The instruction and an example of Task 2-9 Event Extraction.

INSTRUCTION: Below is a sentence from a legal text. Please determine which words in this sentence trigger one or more of the following events: Search/Seizure, Restitution. Provide the words that trigger the events directly, separated by semicolons, for example, "forensic appraisal; fracture". Please answer strictly in this format. Sentence:

QUERY: The public security organ returned the seized vehicles Qiong A·LZ997 and Qiong A·MU397 to Zheng Mouhua and He Moudian, respectively, on August 4, 2014.

ANSWER: seized; returned

Table 18: The instruction and an example of Task 2-10 Trigger Word Extraction.

- ChatLaw (Cui et al., 2023): ChatLaw series contains two model: ChatLaw-13B is fine-tuned based on Ziya-LLaMA-13B-v1 and ChatLaw-33B is fine-tuned based on Anima-33B. Their data primarily consists of forums, news articles, legal provisions, judicial interpretations, legal consultations, law examination questions, and court judgments. Subsequently, it undergoes data cleaning and augmentation processes to construct dialogue datasets.

- LawyerLLaMA (Huang et al., 2023): Lawyer-LLaMA, which is based on the 13-billion parameter Chinese LLaMA model, has been further fine-tuned using the Law SFT dataset to enhance its legal capabilities. They construct a corpus of legal domain Q&A data by having ChatGPT act as a lawyer to generate answers and explanations for questions based on JEC-QA and online legal consultation data. Supervised fine-tuning is performed on these datasets to build large-scale models capable of handling legal tasks.

- FuziMingcha (Wu et al., 2023): FuziMingcha is a legal large language model based on ChatGLM. Its training data can be divided into two main categories: Unsupervised Chinese judicial corpus and judicial supervised tuning data. This wide-ranging content entails not only laws and regulations, judicial interpretations, and verdicts, but also encompasses various high-quality judicial task datasets, such as legal Q&A, similar case retrieval, and legal syllogism. The rich and qualitatively massive training data ensures an accurate and comprehensive coverage of distinctive judicial domain-specific information, thereby solidifying the knowledge foundation of the FuziMingcha judicial model.

- LexiLaw (Lex, 2023): a fine-tuned Chinese legal model based on the ChatGLM-6B with legal datasets. LexiLaw's training data is obtained through a combination of general domain data and legal data.

- WisdomInterrogatory (Wis, 2023): a further pre-trained and fine-tuned model built upon Baichuan-7B. They perform continued pre-trained of this model on legal documents, judicial cases, and legal Q&A data to incorporate legal domain knowledge into the model. Subsequently, the model is supervised fine-tuning on 100k instruction tuning data to endow it with question-answering capabilities. The sources of instruction tuning data include legal documents, judicial cases, and legal Q&A data, where data from the legal field constitutes 30% of the total data, with the remaining being a general dataset.

INSTRUCTION: Provide the relevant articles in the criminal code based on the following facts and charges. Please only include the criminal code article numbers, and place your answers between [Article] and <eoa>. For example, [Article] Article 128 of the Criminal Code, Article 341 of the Criminal Code <eoa>.

QUERY: Fact: The People's Procuratorate of Gongzhuling City, Jilin Province, alleges that on June 18, 2014, the defendant, Mr. Zhang, applied for a 'Da Jilin' credit card at the China Bank Gongzhuling City Branch. As of August 10, 2014, the outstanding balance reached 14,974.52 yuan. Despite two collection attempts by the issuing bank's staff, the defendant still did not repay the amount for over three months.
Charge: Fraud.

ANSWER: Article: Article 196 of the Criminal Code

Table 19: The instruction and an example of Task 3-1 Fact-based Article Prediction.

INSTRUCTION: Please provide legal justification based on specific scenarios and questions, presenting only the relevant legal articles. Each scenario should be associated with a single legal article.

QUERY: Scenario: A company's board of directors has decided to appoint one of its directors as a concurrent manager to enhance the company's management. According to which legal provision, can the board of directors make the decision to have a board member serve concurrently as a manager?

ANSWER: According to Article 114 of the Company Law, the board of directors of a company can decide to appoint one of its members as a concurrent manager.

Table 20: The instruction and an example of Task 3-2 Scene-based Article Prediction.

I Additional Results of ChatGPT and GPT-4

In this section, we present the results of ChatGPT and GPT-4 when we set the temperature to 0. Table 27 shows the results of this experiment.

From these results, it can be observed that setting the temperature to 0.0 results in relatively minor fluctuations compared to the results obtained when the temperature is set to 0.7. Perhaps most importantly, these minor variations in performance numbers do not impact the analyses and conclusions within the text at all.

J Impact of Maximum Context Length

We set the input token length limit to 2048 and an output token length to 1024. All the models we tested can operate normally under this setting. We have statistics on the number of entries in our test dataset where the combined length of each task's Instruction and Question exceeds 2048 characters. The statistical information is shown in Table 28.

From the results in this table, we can observe that the length we have set is quite appropriate, as it successfully covers the vast majority of our test data, including 99.18% of the zero-shot dataset and 99.11% of the one-shot dataset.

K Details of Results

The more detailed results of these models are presented in Tables 29 to Table 36. The performance of most models on zero-shot and one-shot tasks is similar. Unless otherwise specified, we refer to issues that exist in both the zero-shot and one-shot scenarios. We conduct a detailed analysis of certain models that demonstrate notable performance and differences from other models or similar types in each task. We carry out our interpretation through an examination of their output.

We discover that the performance of the LLaMA series models on this test set was the most inferior, even the 70B model underperformed the 7B models in the Chinese oriented LLMs such as the InternLM2-7B-Chat. Though the LLaMA series models are multilingual, they typically generate responses using a substantial amount of English to answer Chinese questions or intersperse English with Chinese in their responses. This leads to their lowest overall scores in evaluations.

In task 1-1, the ChatLaw-33B model scores lower than other large legal language models due to its lack of specific legal provisions knowledge and its repetitive usage of the same legal content to answer different questions.

In task 1-2, the Fuzi-mingcha yielded the lowest scores in the legal model under a zero-shot scenario, principally because it failed to accurately comprehend the question and to present the answer in the format given in the question. Although the task is a multiple choice endeavour, the Fuzi-mingcha tends to answer the question from the stem on its own, rather than selecting the correct answer from the

INSTRUCTION: Please simulate a judge and provide the charge based on the following facts. Only the name of the charge is required. Place your answer between [Charge] and <eoa>. For example, [Charge] Theft; Fraud<eoa>. Please strictly adhere to this format in your response.

QUERY: Facts: The People's Procuratorate of Haidian District, Beijing, in the indictment, alleges the following:
On June 27, 2013, the defendant, Ren (owner of an East Wind Nissan Yida sedan with license plate number Jing PNB057 and insured under this vehicle), in Haidian District of Beijing, fabricated a fictitious traffic accident involving an East Wind Nissan Yida sedan (license plate number Jing PNB057) and a BMW sedan (license plate number Jing QSU596) to deceive China People's Property Insurance Co., Ltd., Beijing Branch, into paying RMB 14,160.8.
On August 15, 2013, the defendant, Ren, in Haidian District of Beijing, fabricated a fictitious traffic accident involving a Jinbei sedan (license plate number Jing KP2320) and a BMW sedan (license plate number Jing QSU596) to deceive China People's Property Insurance Co., Ltd., Beijing Branch, into paying RMB 10,231.2.
On September 26, 2013, the defendant, Ren, was summoned by the public security authorities. All the funds involved in the case were refunded to the victim company after the incident.

ANSWER: Charge: Insurance fraud

Table 21: The instruction and an example of Task 3-3 Charge Prediction.

INSTRUCTION: Based on the following facts, charges, and articles of the criminal code, predict the length of the sentence. Only provide the sentence length in months, and place your answer between [Sentence] and <eoa>. For example, [Sentence] 12 months <eoa>.

QUERY: Facts: The prosecuting authority alleges that on December 23, 2015, the defendant, Yang, took advantage of his position as the former secretary of Qiansuo Street in Jiaojiang District, Taizhou City, and misappropriated RMB 140,000 of pre-collected construction fees from Shangxu Village residents. He used this money to cover expenses related to his eyeglass factory.
On March 20, 2016, the defendant, Yang, once again took advantage of his position and misappropriated RMB 51,765 of Shangxu Village's funds. He used this money to repay a bank loan for his eyeglass factory.
On March 31 of the same year, the defendant, Yang, returned all the misappropriated funds to the village collective account. On February 23, 2017, at approximately 10 o'clock, Yang was summoned to the case by the Economic Investigation Brigade of Jiaojiang Branch, Taizhou Public Security Bureau, at his residence at No. 2 Shangxu Village, Qiansuo Street, Jiaojiang District, Taizhou City.
Charge: Embezzlement

ANSWER: Prison term: 10 months

Table 22: The instruction and an example of Task 3-4 Prison Term Prediction w.o Article.

choices provided.

In task 2-1, the majority of models failed to accurately identify the target for revision in the sentence. Although they output revised sentences as required by the task, they did not accurately correct the errors in the sentences. Wisdom-Interrogatory significantly outperformed other legal models on this task because its training dataset contained a similar task. We posit that training on this task substantially improves the models' ability to accurately identify and correct errors in sentences.

In task 2-2, the performance of Fuzi-Mingcha under the one-shot scenario was significantly lower compared to that in the zero-shot scenario due to its inability to appropriately utilize the provided answer examples for reference despite being aware of their necessity. This failure to correctly draw upon and align with the examples led to notably lower scores for the model.

In task 2-3, Lawyer-LLaMA performs significantly worse in the one-shot scenario than other models and its own performance in the zero-shot scenario. We observe that it is more prone to misunderstanding the intent of the instruction in the one-shot scenario, responding to multi-choice questions as if they are legal advice instructions.

In task 2-4, we observe that Lawyer-LLaMA significantly underperforms compared to other legal models due to its misinterpretation of the task directions, which it erroneously perceived as consultative questions, subsequently providing incorrect answers.

In tasks 2-5, Fuzi-Mingcha significantly outperformed other models. This superior performance can be attributed to the inclusion of similar tasks within its supervised fine-tuning dataset, which evidently enhanced its ability to excel in these specific tasks.

In task 2-6, InternLM2-20B-Chat exhibits significantly lower scores compared to other models due to its tendency to generate an excessive amount of irrelevant content when presenting entities. Additionally, the model's responses did not adhere to the conciseness required by the prompts.

INSTRUCTION: Based on the following facts, charges, and articles of the criminal code, predict the length of the sentence. Only provide the sentence length in months, and place your answer between [Sentence] and <eoa>. For example, [Sentence] 12 months <eoa>.

QUERY: Facts: The public prosecution accuses that on July 3, 2014, at around 12 o'clock, the defendant, Sun, became very angry when he found that a fellow villager, Zhang, had not returned home after going out due to drinking. He then climbed over the wall into Zhang's courtyard. Inside the house, he smashed and tore apart Zhang's TV, stereo, clothes, and other belongings, burning some of the clothes. When Zhang returned home and tried to stop the defendant, Sun, he was punched in the face by the defendant. According to the assessment, the damaged property was worth 1,580 yuan. On October 10, 2014, at around 8 pm, the defendant, Sun, sent a message to Zhang, threatening to jump into his house. Zhang hastily called his brother-in-law, and upon learning of the situation, his brother-in-law, Sun, reported it to the police. After the incident, both parties reached a compensation agreement. The defendant, Sun, compensated the victim, Zhang, for his losses and obtained the victim's forgiveness.
Charge: Illegal intrusion into a residence.
Legal Article: Article 245 of the Criminal Law.
Legal article content: Article 245 - Whoever illegally searches another person's body or residence, or illegally intrudes into another person's residence, shall be sentenced to fixed-term imprisonment of not more than three years or criminal detention. If judicial personnel abuse their power and commit the crime mentioned in the preceding paragraph, they shall be punished more severely.

ANSWER: Prison term: 4 months

Table 23: The instruction and an example of Task 3-5 Prison Term Prediction w. Article.

INSTRUCTION: Please analyze the following case using legal knowledge and select the correct answer from A, B, C, D, and write it between [Correct Answer] and <eoa>. For example, [Correct Answer] A <eoa>. Please adhere strictly to this format in your response.

QUERY: During the May Day holiday, tourist A had a conflict with tourist B over the purchase of tickets to the Potala Palace in Tibet and injured tourist B. Upon investigation, it was found that tourist A, who is of Han ethnicity, is from Guangdong province but has been living in Fujian for several years and has some knowledge of Tibetan. In this case, what language should the Lhasa police use to interrogate tourist A? A: Mandarin; B: Fujianese; C: Cantonese; D: Tibetan.

ANSWER: Correct answer: D.

Table 24: The instruction and an example of Task 3-6 Case Analysis.

In task 2-7, Ziya-LLaMA-13B demonstrated notably inferior performance compared to other models, primarily due to its inability to comprehend the meaning of the questions. Consequently, it generated outputs that included nonsensical characters, such as ':'. This deficiency led to lower scores and a higher rate of abstention for Ziya-LLaMA-13B on this task.

In task 2-8, we observe that the performance of Wisdom-Interrogatory and Lawyer-LLaMA under one-shot scenario was notably inferior to their performance in the zero-shot scenario. Analysis of the responses generated by these models reveals that during one-shot scenarios, both Wisdom-Interrogatory and Lawyer-LLaMA tend to replicate the examples provided to them, resulting in suboptimal outcomes.

In task 2-9, the ChatGLM model does not accurately comprehend the intended meaning of the prompts. It fails to recognize that the task requires the identification of relevant events from provided scenarios. Instead, the model treats it as an inquiry task, generating responses that are unrelated to the question. Consequently, its performance on these tasks is significantly inferior to that of other comparable models.

In task 2-10, we observe that the legal models scored significantly lower than the Chinese oriented models because they failed to recognize the trigger words and did not understand the intent of the question. The responses they generated were mostly irrelevant to the question, resulting in low scores.

In task 3-1, both Lawyer-LLaMA and ChatLaw-33B demonstrate significant difficulty in comprehending the implications of the posed questions. Their responses were unrelated to the issues, which resulted in their performance being markedly lower than that of other law specific large language models on this task.

In task 3-2, we observe a significantly lower performance of the InternLM2 series in comparison to other Chinese Oriented LLMs. This can be attributed to their inability to comprehend that the essence of the task was to deliver relevant laws. Instead, these models analyzed and answered the questions embedded in the task.

In task 3-3, the significant rates of omission observed in most models can be attributed to the

INSTRUCTION: Please calculate the total amount of the crimes mentioned in the document carefully. You don't need to provide the calculation process; just provide the final amount between [Amount] and <eoa>. For example, [Amount] 2000 yuan <eoa>.

QUERY: Document: Upon trial and investigation, it was ascertained that in the early hours of September 23, 2016, the defendant, Liang, drove a motorcycle to the entrance of Hutou Village in Pingma Village Committee, Xiaojiang Street, Pubei County, Guangxi. He stole 11 castrated chickens and 12 hens from the victim, Huang 1, as well as 2 castrated chickens from Huang 2 and 1 hen from Ning. The defendant was apprehended by Huang 1 and others while fleeing the scene and subsequently reported to the public security authorities. According to the appraisal conducted by the Price Certification Center of Pubei County Price Bureau, the total market value of the stolen chickens amounted to RMB 2,094.

ANSWER: The amount involved in the above-mentioned crime: 2094.0 yuan.

Table 25: The instruction and an example of Task 3-7 Crime Amount Calculation.

INSTRUCTION: Please answer the following questions, first provide the answer, and then provide the corresponding legal basis:

QUERY: Is the accident determination related to the driver crossing the centerline and hitting a person, resulting in their death despite rescue efforts? How long should it take to issue the accident determination report?

ANSWER: Answer: In the case of a traffic accident resulting in a fatality, the traffic management department of the public security organ should issue the Road Traffic Accident Determination Report within ten days from the date of on-site investigation. For hit-and-run traffic accident cases, the Road Traffic Accident Determination Report should be issued within ten days after the apprehension of the hit-and-run vehicle and driver.
Legal Basis: Article 62 of the 'Provisions on the Procedures for Handling Road Traffic Accidents' states that the traffic management department of the public security organ should issue the Road Traffic Accident Determination Report within ten days from the date of on-site investigation in cases of traffic accidents resulting in fatalities. In hit-and-run cases, the Road Traffic Accident Determination Report should be issued within ten days after the apprehension of the hit-and-run vehicle and driver. In cases requiring examination or appraisal, the Road Traffic Accident Determination Report should be issued within five days from the date of determination based on the examination report or appraisal opinion.

Table 26: The instruction and an example of Task 3-8 Consultation.

	ChatGPT	ChatGPT	GPT4	GPT4
Task ID	0-Shot	1-Shot	0-Shot	1-Shot
1-1	16.13	16.65	15.68	17.72
1-2	36.00	36.60	54.60	54.60
2-1	11.98	15.49	14.51	19.77
2-2	39.60	41.20	44.60	44.40
2-3	54.01	54.69	70.54	71.30
2-4	41.60	40.60	45.20	43.60
2-5	54.28	62.44	56.62	65.24
2-6	70.77	75.63	76.78	80.44
2-7	33.89	42.42	39.86	41.82
2-8	37.20	40.40	60.60	60.80
2-9	68.25	68.82	79.19	77.46
2-10	40.64	42.76	66.56	66.24
3-1	33.79	33.19	54.09	54.90
3-2	33.71	34.75	28.13	33.47
3-3	38.44	38.03	42.55	41.21
3-4	81.54	76.74	82.64	83.48
3-5	77.06	74.35	81.53	82.60
3-6	29.40	32.40	49.40	51.00
3-7	63.20	67.60	78.60	79.20
3-8	17.65	17.27	20.11	20.04

Table 27: Results of ChatGPT and GPT-4 when the temperature is set to 0.

Task ID	zero-shot	one-shot
1-1	0	0
1-2	0	0
2-1	0	0
2-2	15	17
2-3	0	0
2-4	0	0
2-5	0	1
2-6	0	0
2-7	0	0
2-8	0	0
2-9	0	0
2-10	0	0
3-1	14	14
3-2	0	0
3-3	8	9
3-4	14	15
3-5	30	32
3-6	0	0
3-7	1	1
3-8	0	0

Table 28: Number of documents that are longer than the maximum input token length for each task and each evaluation setting.

necessity for the model to be aware of relevant charges in criminal law and generate accurate charges accordingly. However, the charges generated by these models predominantly deviate from the correct answers, resulting in high rates of abstention.

In task 3-4, under the scenarios of zero-shot and one-shot, Ziya-LLaMA-13B significantly lags behind other models of its kind due to its inability to comprehend the task, leading to the generation of some meaningless characters such as '.'.

In task 3-5, under the scenarios of zero-shot and one-shot, a significant portion of the content generated by Ziya-LLaMA-13B is blank, causing its outcomes to be notably inferior to other models, along with a high abstention rate. Meanwhile, the ChatLaw-33B model generates a high amount of irrelevant responses during its reply, resulting in a significantly higher abstention rate compared to other legal models.

In task 3-6, Lawyer-LLaMA significantly underperformed compared to other legal models in a One-Shot scenario, exhibiting a high rate of omissions. This was primarily due to its inability to comprehend the task instructions, which required selecting answers from given options. Instead, it independently analyzed the questions in the prompts and provided answers. Similarly, in a zero-shot context, Fuzi-Mingcha also failed to understand the task requirements, generating content irrelevant to the question, resulting in it achieving the lowest score among all the legal models.

In task 3-7, larger-scale models such as GPT-4 and Qwen1.5-72B-Chat exhibited superior performance in numerical computations, significantly outperforming other models. Furthermore, ChatGLM2-6B and Ziya-LLaMA-13B scored significantly higher in one-shot scenarios compared to zero-shot scenarios. We observed that these two models failed to accurately comprehend the essence of the questions in zero-shot scenarios. Ziya-LLaMA-13B tended to paraphrase the questions in its responses rather than delivering actual answers. On the other hand, ChatGLM2-6B simulated a judge assigning a fine instead of providing a calculation for the amount involved in the case.

In task 3-8, Wisdom-Interrogatory considerably surpasses other legal models. This shows its superior capability of adhering to the provided format illustrated in the question while generating responses. It presents the answer initially and then cites the relevant law as a reference.

L Results Grouped by Models' Size

The performance correlates strongly with the parameter count. We present tables and plots where models of similar size are grouped to display the results. Figure 2 shows the overall zero-shot / one-shot performance of each model, where the models are ranked by their size and scores obtained by averaging their scores over the 20 tasks. We observe that larger models usually outperform smaller models within the same family. For example, in the Qwen series, the Qwen-1.5-72B model outperforms the Qwen-1.5-14B and Qwen-1.5-7B models in zero shot setting on average performance. However, there are some outliers. Notably, InternLM2-20B does not perform better on average than InternLM2-7B. Among models with similar parameters, some general-purpose models still outperform several large language models tailored for the legal domain. For example, InternLM2-7B outperforms Wisdom-7B in both zero-shot and one-shot settings. This indicates that legal specific LLMs still have a lot of room for improvement.

M Details of Dataset Statistics

We present detailed statistical information of our dataset in Table 37 to Table 40.

model_name	1-1	1-2	Memorization
≈7B			
ChatGLM2-6B	15.37	10.60	12.98
LexiLaw	16.96	21.00	18.98
Fuzi-Mingcha	25.22	7.80	16.51
LLaMA-2-Chat-7B	0.80	6.40	3.60
Wisdom-Interrogatory	43.05	15.40	29.23
Baichuan2-7B-Chat	18.37	36.40	27.38
Qwen1.5-7B-Chat	18.80	51.00	34.90
InternLM2-7B-Chat	13.03	50.20	31.61
≈ 13B			
LLaMA-2-Chat-13B	1.05	3.20	2.12
Ziya-LLaMA-13B	15.04	30.00	22.52
Lawyer-LLaMA	12.33	23.20	17.77
ChatLaw-13B	14.85	28.40	21.63
Baichuan2-13B-Chat	20.16	41.40	30.78
Qwen1.5-14B-Chat	22.30	60.40	41.35
20B			
InternLM2-20B-Chat	11.95	42.00	26.98
33B			
ChatLaw-33B	11.74	28.60	20.17
≈ 70B			
LLaMA-2-Chat-70B	0.87	10.00	5.43
StableBeluga2	14.58	34.60	24.59
Qwen1.5-72B-Chat	29.13	76.40	52.77
Commercial models			
ChatGPT	15.86	36.00	25.93
GPT-4	15.38	55.20	35.29

%abstention 0% 25% 50% 75% 100%

LLMs Category Multilingual Chinese Oriented Law Specific

Table 29: Zero-shot Results on Legal Knowledge Memorization Tasks

model_name	2-1	2-2	2-3	2-4	2-5	2-6	2-7	2-8	2-9	2-10	Understanding
≈7B											
ChatGLM2-6B	1.84	1.65	13.20	22.40	30.33	36.81	19.50	4.00	11.60	2.81	14.41
LexiLaw	6.24	3.30	15.60	22.80	45.39	48.74	33.12	21.60	15.30	11.17	22.32
Fuzi-Mingcha	4.93	19.59	28.46	18.60	97.59	44.07	54.32	8.80	16.90	7.78	30.10
LLaMA-2-Chat-7B	0.25	5.36	16.92	5.20	4.90	5.62	1.13	6.20	14.77	2.06	6.24
Wisdom-Interrogatory	30.97	7.84	36.72	21.00	35.56	57.06	33.34	10.60	15.98	6.24	25.53
Baichuan2-7B-Chat	27.17	14.60	41.11	21.20	59.17	42.72	34.26	31.00	35.86	24.04	33.11
Qwen1.5-7B-Chat	12.00	31.80	46.86	39.20	62.57	20.83	30.59	38.40	58.65	16.37	35.73
InternLM2-7B-Chat	36.78	39.20	54.52	43.80	47.21	51.50	33.60	43.20	63.89	36.32	45.00
≈ 13B											
LLaMA-2-Chat-13B	0.84	9.48	40.80	24.80	6.08	2.20	1.16	9.20	35.74	4.25	13.46
Ziya-LLaMA-13B	8.13	2.06	44.18	24.60	27.93	43.44	2.00	11.80	31.82	3.06	19.90
Lawyer-LLaMA	4.33	8.25	15.88	4.40	34.61	41.65	38.51	9.60	29.78	2.38	18.94
ChatLaw-13B	12.22	2.68	42.24	27.60	39.11	54.89	38.45	18.60	31.74	14.56	28.21
Baichuan2-13B-Chat	32.49	35.40	40.38	31.20	60.43	65.33	36.37	36.20	43.23	29.89	41.09
Qwen1.5-14B-Chat	19.84	40.60	45.48	40.80	59.43	47.35	33.94	42.20	62.23	27.10	41.90
20B											
InternLM2-20B-Chat	33.19	43.20	54.95	44.20	33.45	12.51	34.65	11.00	62.83	31.44	36.14
33B											
ChatLaw-33B	3.67	8.04	32.08	19.80	37.16	30.14	35.47	26.40	22.14	10.56	22.55
≈ 70B											
LLaMA-2-Chat-70B	0.74	7.22	33.38	30.40	5.22	34.72	1.25	8.40	39.75	12.61	17.37
StableBeluga2	7.70	25.57	44.20	39.00	52.03	65.54	39.07	45.80	65.27	41.64	42.58
Qwen1.5-72B-Chat	35.01	48.60	62.05	39.00	66.47	75.53	34.81	54.40	70.55	43.29	52.97
Commercial models											
ChatGPT	9.10	32.37	51.73	41.20	53.75	69.55	33.49	36.40	66.48	39.05	43.31
GPT-4	12.53	41.65	69.79	44.00	56.50	76.60	37.92	61.20	78.82	65.09	54.41

%abstention 0% 25% 50% 75% 100%

LLMs Category Multilingual Chinese Oriented Law Specific

Table 30: Zero-shot Results on Legal Knowledge Understanding Tasks

model_name	3-1	3-2	3-3	3-4	3-5	3-6	3-7	3-8	Application
≈7B									
ChatGLM2-6B	23.12	24.96	27.12	59.79	72.84	10.80	16.80	17.56	31.62
LexiLaw	13.15	35.78	39.99	78.08	74.92	20.80	35.80	15.82	39.29
Fuzi-Mingcha	25.19	22.18	55.93	77.23	75.52	7.00	47.20	16.64	40.86
LLaMA-2-Chat-7B	1.87	0.64	11.57	10.13	7.21	5.60	36.20	0.36	9.20
Wisdom-Interrogatory	32.84	32.01	35.09	80.36	81.10	15.40	17.40	20.17	39.29
Baichuan2-7B-Chat	54.86	25.52	46.63	78.15	71.89	35.00	51.00	18.47	47.69
Qwen1.5-7B-Chat	57.53	31.93	45.35	79.26	79.53	45.00	43.00	19.51	50.14
InternLM2-7B-Chat	63.79	14.12	48.91	81.42	80.11	39.60	55.40	19.32	50.33
≈ 13B									
LLaMA-2-Chat-13B	1.40	0.49	15.29	50.57	52.64	4.60	28.60	0.40	19.25
Ziya-LLaMA-13B	10.20	21.23	25.93	0.97	0.57	30.60	12.80	13.36	14.46
Lawyer-LLaMA	0.60	25.94	31.30	74.19	75.52	17.80	39.20	16.94	35.19
ChatLaw-13B	33.28	31.55	27.90	76.18	73.57	28.80	41.40	17.17	41.23
Baichuan2-13B-Chat	69.31	28.20	52.14	77.20	78.47	40.60	52.80	20.20	52.37
Qwen1.5-14B-Chat	67.24	26.83	47.81	77.47	74.76	56.60	61.60	22.98	54.41
20B									
InternLM2-20B-Chat	69.98	12.96	48.00	81.81	83.19	17.40	63.20	17.45	49.25
33B									
ChatLaw-33B	5.35	26.02	12.73	67.00	53.63	34.20	41.60	16.55	32.14
≈ 70B									
LLaMA-2-Chat-70B	6.82	0.89	14.91	17.54	12.81	9.60	42.40	4.61	13.70
StableBeluga2	16.41	24.52	22.82	76.06	65.35	34.40	56.60	13.39	38.69
Qwen1.5-72B-Chat	72.42	29.67	57.07	81.32	79.95	70.40	74.80	24.30	61.24
Commercial models									
ChatGPT	29.50	31.30	35.52	78.75	76.84	27.40	61.20	17.45	44.74
GPT-4	52.47	27.54	41.99	82.62	81.91	48.60	77.60	19.65	54.05

%abstention 0% 25% 50% 75% 100%

LLMs Category Multilingual Chinese Oriented Law Specific

Table 31: Zero-shot Results on Legal Knowledge Application Tasks

model_name	Memorization	Understanding	Application	Overall
≈7B				
ChatGLM2-6B	12.98	14.41	31.62	21.15
LexiLaw	18.98	22.32	39.29	28.78
Fuzi-Mingcha	16.51	30.10	40.86	33.05
LLaMA-2-Chat-7B	3.60	6.24	9.20	7.16
Wisdom-Interrogatory	29.23	25.53	39.29	31.41
Baichuan2-7B-Chat	27.38	33.11	47.69	38.37
Qwen1.5-7B-Chat	34.90	35.73	50.14	41.41
InternLM2-7B-Chat	31.61	45.00	50.33	45.80
≈ 13B				
LLaMA-2-Chat-13B	2.12	13.46	19.25	14.64
Ziya-LLaMA-13B	22.52	19.90	14.46	17.99
Lawyer-LLaMA	17.77	18.94	35.19	25.32
ChatLaw-13B	21.63	28.21	41.23	32.76
Baichuan2-13B-Chat	30.78	41.09	52.37	44.57
Qwen1.5-14B-Chat	41.35	41.90	54.41	46.85
20B				
InternLM2-20B-Chat	26.98	36.14	49.25	40.47
33B				
ChatLaw-33B	20.17	22.55	32.14	26.14
≈ 70B				
LLaMA-2-Chat-70B	5.43	17.37	13.70	14.71
StableBeluga2	24.59	42.58	38.69	39.23
Qwen1.5-72B-Chat	52.77	52.97	61.24	56.26
Commercial models				
ChatGPT	25.93	43.31	44.74	42.15
GPT-4	35.29	54.41	54.05	52.35

%abstention 0% 25% 50% 75% 100%

LLMs Category Multilingual Chinese Oriented Law Specific

Table 32: Zero-shot Results on Overall

model_name	1-1	1-2	Memorization
≈7B			
ChatGLM2-6B	14.67	24.80	19.74
LexiLaw	15.47	14.40	14.94
Fuzi-Mingcha	20.21	12.80	16.50
LLaMA-2-Chat-7B	0.82	7.40	4.11
Wisdom-Interrogatory	37.41	15.20	26.30
Baichuan2-7B-Chat	17.86	37.20	27.53
Qwen1.5-7B-Chat	18.15	46.20	32.18
InternLM2-7B-Chat	17.04	47.00	32.02
≈ 13B			
LLaMA-2-Chat-13B	1.62	3.80	2.71
Ziya-LLaMA-13B	0.00	28.00	14.00
Lawyer-LLaMA	13.04	10.60	11.82
ChatLaw-13B	15.98	29.40	22.69
Baichuan2-13B-Chat	21.01	41.20	31.11
Qwen1.5-14B-Chat	20.84	58.80	39.82
20B			
InternLM2-20B-Chat	21.07	50.60	35.84
33B			
ChatLaw-33B	14.36	27.80	21.08
≈ 70B			
LLaMA-2-Chat-70B	0.73	11.00	5.86
StableBeluga2	15.03	36.00	25.51
Qwen1.5-72B-Chat	25.71	74.40	50.06
Commercial models			
ChatGPT	16.15	37.20	26.67
GPT-4	17.21	54.80	36.00

%abstention	0%	25%	50%	75%	100%
LLMs Category	Multilingual	Chinese Oriented	Law Specific		

Table 33: One-shot Results on Legal Knowledge Memorization Tasks

model_name	2-1	2-2	2-3	2-4	2-5	2-6	2-7	2-8	2-9	2-10	Understanding
≈7B											
ChatGLM2-6B	3.17	10.80	10.31	4.80	38.38	36.61	15.59	12.80	26.72	5.64	16.48
LexiLaw	4.18	15.40	21.49	27.20	41.64	31.54	34.57	6.00	19.84	8.36	21.02
Fuzi-Mingcha	2.86	2.40	17.44	8.80	93.35	42.28	31.43	11.40	21.26	7.04	23.83
LLaMA-2-Chat-7B	0.13	3.40	16.78	5.40	0.64	7.87	1.60	4.20	16.56	2.29	5.89
Wisdom-Interrogatory	22.16	10.00	24.29	13.40	13.23	40.10	39.71	0.20	15.81	4.02	18.29
Baichuan2-7B-Chat	27.17	34.00	43.69	25.60	59.35	51.82	39.69	22.20	27.59	20.67	35.18
Qwen1.5-7B-Chat	14.51	22.80	51.29	40.00	64.60	61.40	33.47	39.00	62.96	22.41	41.24
InternLM2-7B-Chat	36.78	40.00	49.55	41.80	61.62	64.95	37.12	44.80	66.54	40.18	48.33
≈ 13B											
LLaMA-2-Chat-13B	0.23	6.20	31.00	17.80	0.97	3.16	1.03	5.20	26.85	2.83	9.53
Ziya-LLaMA-13B	10.27	15.40	42.75	25.20	7.64	56.84	2.75	17.00	31.13	2.65	21.16
Lawyer-LLaMA	4.90	19.20	9.03	3.00	39.65	36.33	37.10	0.40	33.19	6.12	18.89
ChatLaw-13B	13.01	9.00	30.91	26.60	41.41	60.68	42.71	20.20	40.27	17.37	30.22
Baichuan2-13B-Chat	38.43	35.40	50.22	29.80	65.66	65.59	43.20	32.60	47.90	28.73	43.75
Qwen1.5-14B-Chat	19.84	35.40	48.38	40.40	71.18	64.28	38.66	47.00	68.68	30.34	46.42
20B											
InternLM2-20B-Chat	46.81	43.40	54.65	43.20	70.72	27.05	37.90	22.60	67.69	42.85	45.69
33B											
ChatLaw-33B	4.30	12.20	33.49	4.20	38.87	28.83	34.20	15.40	26.18	15.95	21.36
≈ 70B											
LLaMA-2-Chat-70B	0.00	17.40	38.05	28.40	1.16	3.13	1.53	6.40	34.88	8.39	13.93
StableBeluga2	8.93	15.00	41.76	38.00	53.55	64.99	45.06	37.60	65.89	40.54	41.13
Qwen1.5-72B-Chat	35.01	44.20	65.35	40.60	78.46	73.83	42.11	57.60	74.71	37.35	54.92
Commercial models											
ChatGPT	13.50	40.60	54.01	41.40	61.98	74.04	40.68	37.40	67.59	40.04	47.12
GPT-4	18.31	46.00	69.99	44.40	64.80	79.96	40.52	59.00	76.55	65.26	56.48

%abstention	0%	25%	50%	75%	100%
LLMs Category	Multilingual	Chinese Oriented	Law Specific		

Table 34: One-shot Results on Legal Knowledge Understanding Tasks

model_name	3-1	3-2	3-3	3-4	3-5	3-6	3-7	3-8	Application
≈7B									
ChatGLM2-6B	14.34	34.23	32.09	38.25	61.35	16.60	42.80	14.89	31.82
LexiLaw	15.41	33.94	34.03	73.66	70.93	12.20	33.20	14.68	36.01
Fuzi-Mingcha	3.86	32.96	43.60	78.95	79.00	13.80	38.20	13.95	38.04
LLaMA-2-Chat-7B	1.00	1.22	6.55	17.00	8.16	5.40	29.80	0.32	8.68
Wisdom-Interrogatory	20.02	23.33	39.22	81.16	81.57	13.00	40.40	20.67	39.92
Baichuan2-7B-Chat	59.61	34.58	47.93	78.69	69.29	36.20	51.60	19.44	49.67
Qwen1.5-7B-Chat	53.57	33.86	44.91	80.86	78.02	47.20	42.40	19.84	50.08
InternLM2-7B-Chat	64.15	29.35	51.03	80.11	80.21	41.40	56.60	20.42	52.91
≈ 13B									
LLaMA-2-Chat-13B	0.53	13.32	16.53	33.50	38.01	2.40	35.80	1.31	17.68
Ziya-LLaMA-13B	5.83	33.63	25.98	4.51	0.55	26.00	24.20	15.53	17.03
Lawyer-LLaMA	0.33	27.23	19.36	70.99	73.56	6.60	33.80	16.02	30.99
ChatLaw-13B	25.99	33.96	12.24	74.31	73.01	26.80	42.00	16.72	38.13
Baichuan2-13B-Chat	65.64	34.09	52.41	77.18	78.04	41.40	56.60	22.49	53.48
Qwen1.5-14B-Chat	64.55	34.97	51.41	74.63	76.21	57.60	63.60	21.84	55.60
20B									
InternLM2-20B-Chat	73.85	20.51	51.00	81.57	82.04	32.00	62.60	22.79	53.29
33B									
ChatLaw-33B	4.89	27.97	17.54	63.30	53.03	26.20	43.20	16.29	31.55
≈ 70B									
LLaMA-2-Chat-70B	6.39	1.41	15.87	13.17	6.22	13.00	41.60	4.57	12.78
StableBeluga2	16.87	32.44	23.07	75.80	63.59	33.00	56.00	16.24	39.63
Qwen1.5-72B-Chat	73.79	36.10	60.01	80.77	79.11	68.80	75.00	24.67	62.28
Commercial models									
ChatGPT	30.81	34.49	34.55	77.12	73.72	31.60	66.40	17.17	45.73
GPT-4	53.20	33.15	41.30	83.21	82.74	49.60	77.00	19.90	55.01

%abstention	0%	25%	50%	75%	100%
LLMs Category	Multilingual	Chinese Oriented	Law Specific		

Table 35: One-shot Results on Legal Knowledge Application Tasks

model_name	Memorization	Understanding	Application	Overall
≈7B				
ChatGLM2-6B	19.74	16.48	31.82	22.94
LexiLaw	14.94	21.02	36.01	26.41
Fuzi-Mingcha	16.50	23.83	38.04	28.78
LLaMA-2-Chat-7B	4.11	5.89	8.68	6.83
Wisdom-Interrogatory	26.30	18.29	39.92	27.74
Baichuan2-7B-Chat	27.53	35.18	49.67	40.21
Qwen1.5-7B-Chat	32.18	41.24	50.08	43.87
InternLM2-7B-Chat	32.02	48.33	52.91	48.53
≈ 13B				
LLaMA-2-Chat-13B	2.71	9.53	17.68	12.11
Ziya-LLaMA-13B	14.00	21.16	17.03	18.79
Lawyer-LLaMA	11.82	18.89	30.99	23.02
ChatLaw-13B	22.69	30.22	38.13	32.63
Baichuan2-13B-Chat	31.11	43.75	53.48	46.38
Qwen1.5-14B-Chat	39.82	46.42	55.60	49.43
20B				
InternLM2-20B-Chat	35.84	45.69	53.29	47.74
33B				
ChatLaw-33B	21.08	21.36	31.55	25.41
≈ 70B				
LLaMA-2-Chat-70B	5.86	13.93	12.78	12.67
StableBeluga2	25.51	41.13	39.63	38.97
Qwen1.5-72B-Chat	50.06	54.92	62.28	57.38
Commercial models				
ChatGPT	26.67	47.12	45.73	44.52
GPT-4	36.00	56.48	55.01	53.85

%abstention	0%	25%	50%	75%	100%
LLMs Category	Multilingual	Chinese Oriented	Law Specific		

Table 36: One-shot Results on Overall

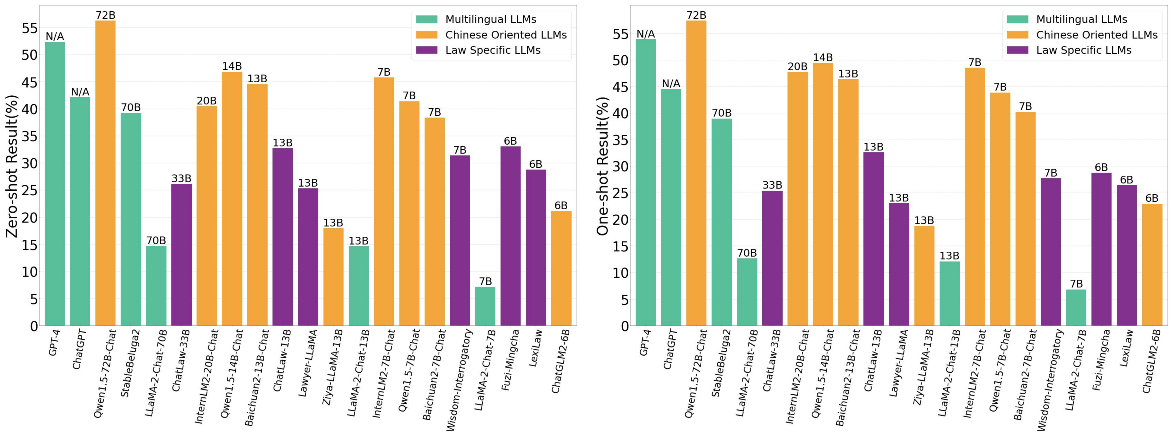

Figure 2: We rank these 21 models based on their size and their average performance under the same size. The left figure shows the models' performance in the zero-shot setting, while the right figure displays their performance in the one-shot setting.

ID	**Instruction**		**Input**			**Output**		
	Zeroshot	**Oneshot**	**Max**	**Min**	**Avg**	**Max**	**Min**	**Avg**
1-1	18	71	36	13	21.742	631	16	103.742
1-2	73	141	429	32	145.41	7	7	7.0
2-1	74	133	452	20	79.614	452	19	79.604
2-2	159	342	4364	156	663.052	16	10	12.534
2-3	214	256	283	10	61.438	48	8	16.006
2-4	174	262	182	6	38.21	4	2	3.85
2-5	21	632	1419	433	734.01	149	6	25.336
2-6	81	122	293	13	69.758	248	0	57.818
2-7	26	556	369	42	323.628	393	27	144.236
2-8	101	389	943	180	382.334	12	12	12.0
2-9	159	208	314	7	55.308	23	2	5.328
2-10	95	146	281	6	62.014	27	1	4.912
3-1	75	182	22142	51	603.968	30	10	11.848
3-2	41	138	154	31	71.206	389	34	127.55
3-3	76	182	9784	26	508.15	52	5	13.082
3-4	68	214	22172	77	615.816	8	5	6.77
3-5	68	214	23421	174	938.584	8	5	6.77
3-6	80	170	858	54	212.434	7	7	7.0
3-7	72	186	1980	63	462.584	22	17	19.386
3-8	28	185	100	50	57.62	731	187	395.02

Table 37: Detailed statistical information for each task, where Instruction represents the length of instructions, Input represents the length of input questions, and Output represents the length of answers.

ID	**Task Type**	**Classes**
1-2	Single-Label Classification	4
2-2	Single-Label Classification	16
2-3	Multi-Label Classification	20
2-4	Single-Label Classification	20
2-6	Extraction	10
2-8	Single-Label Classification	5
2-9	Single-Label Classification	20
3-1	Multi-Label Classification	118
3-3	Multi-Label Classification	372
3-6	Single-Label Classification	4

Table 38: Single/Multi-Label Classification And Extraction Tasks Information. We statistic the number of categories in these tasks.

ID	Task Type	Single Label Number	Multi Label Number
2-3	Multi-Label Classification	252	248
3-1	Multi-Label Classification	339	161
3-3	Multi-Label Classification	336	164

Table 39: Multi-Label Classification Tasks information. We statistic the number of single-label instances and the number of multi-label instances.

ID	Task Type	Max Trigger Word	Min Trigger Word	Avg Trigger Word
2-3	Extraction	8	1	1.914

Table 40: Trigger word extraction tasks information. We statistic the max number, the min number, and the average number of trigger word in this task.