CoLLaM: A Comprehensive Benchmark for Evaluating Large Language Models in Legal Domain

Anonymous ACL submission

Abstract

Large language models (LLMs) have made significant progress in natural language processing tasks and have shown considerable potential in the legal domain. However, the legal applications often have high requirements on accuracy, reliability and fairness. Applying existing LLMs to legal systems without careful evaluation of their potentials and limitations could lead to significant risks in legal practice. Therefore, to facilitate the healthy development and application of LLMs in the legal domain, we propose a comprehensive benchmark CoLLaM for evaluating LLMs in legal domain. Specifically, CoLLaM is developed based on the language abilities of LLMs and the practical requirements of the legal domain. It introduces a new legal cognitive ability taxonomy (LCAT) featuring six distinctive levels: Memorization, Understanding, Logic Inference, Discrimination, Generation, and Ethic. Leveraging this taxonomy, we collected 13,650 questions across 23 tasks and evaluated them against 38 open-source and commercial LLMs. Our experimental results led to interesting findings and indicate that applying LLMs in the legal domain still has a long way to go. The details of CoLLaM can be found on the anonymous website `https://anonymous.4open.science/r/CoLLaM-31F2`.

1 Introduction

Recently, the rapid development of large language models (LLMs) has brought new opportunities to the research of general artificial intelligence. A series of models (e.g., ChatGPT), with their extensive knowledge and outstanding language processing ability, have demonstrated excellent performance in various language processing tasks such as text generation, machine translation, and dialogue systems (Chung et al., 2022; Brown et al., 2020; Chen et al., 2021; Wei et al., 2022; Peng et al., 2023). Meanwhile, LLMs have profoundly impacted the work patterns of legal practitioners and the development of the legal field. Recent studies show that the GPT-4 has the ability to pass the U.S. Judicial Exam (Katz et al., 2023). By interacting with large language models, lawyers and judges can analyze legal documents more efficiently, obtaining comprehensive and valuable information and judicial advice. This has led to a growing trend among legal practitioners to incorporate LLMs as a vital supportive instrument in legal proceedings(Cui et al., 2023a; Savelka et al., 2023b).

Despite the considerable potential of large language models, there are still concerns about their application in the legal domain (Savelka et al., 2023a; Nay et al., 2023). In contrast to the decision-making of human, which is grounded in professional knowledge and logical reasoning, LLMs derive decisions from patterns and connections extracted from massive amounts of training data. Therefore, these models, predicated on probabilistic frameworks, often fall short of ensuring the reliability and explainability of their output (Floridi and Chiriatti, 2020). Given the judicial system is an essential component of social stability and demands a high level of expertise and precision, the deployment of LLMs in legal domain raises concerns. Substandard legal texts or flawed judicial guidance may mislead legal practitioners and increase their workload. More seriously, it may undermine the effectiveness and fairness of judicial proceedings and judgments, bringing considerable systemic risks.

The great potential and risks of LLMs in the legal domain give rise to the urgent need for professional performance benchmark (Sun, 2023). Although numerous methods for evaluating the abilities of LLMs have been developed, most focus on assessing their generalist abilities on non-professional or semi-professional texts. These assessments provide limited guidance for highly specialized fields such as the legal domain.(Zhong et al., 2023; Huang et al., 2023b; Chalkidis et al., 2021). For instance, the well-known Chinese language model evaluation framework, C-Eval (Huang et al., 2023b), primarily uses test questions from high school and university courses. However, in judicial applications, tasks like case summarization, legal case retrieval, and judgment prediction require LLMs to consider precise legal knowledge and complex legal contexts. The content often involves highly specialized elements such as judicial interpretation and reasoning. To the best of our knowledge, existing general evaluation benchmarks are unable to reflect or capture the complexity of judicial cognition and decision-making. How to assess the potential and inherent limitations of LLMs in legal domain is still an open question.

To accurately evaluate the ability of large language models in the legal domain, we construct CoLLaM. CoLLaM focuses on practical legal applications under the Chinese legal system, involving how legal professionals manage, contemplate, and resolve legal issues. Specifically, we first propose a new legal cognitive ability taxonomy (LCAT) that consists of six levels: Memorization, Understanding, Logic Inference, Discrimination, Generation, and Ethic. Based on these cognitive levels, we collect 13,650 questions covering 23 legal tasks. To our knowledge, CoLLaM is the largest and most comprehensive Chinese legal benchmarking dataset for evaluating LLMs. We conduct a thorough evaluation of 38 popular LLMs, including General LLMs and Legal-specific LLMs. The experimental results show that the existing LLMs are ineffective and unreliable in addressing legal problems. We expect more participation and contribution to facilitate the development of more advanced legal LLMs.

2 Related Work

In recent years, large language models (LLMs) have drawn great attention in academia and industry for their excellent performance and wide applicability (OpenAI, 2023; Zeng et al., 2022). Models such as ChatGPT and ChatGLM achieve excellent performance across various tasks through mechanisms such as pre-training, supervised fine-tuning, and alignment with human or AI feedback (Bai et al., 2022; Christiano et al., 2017). By learning from massive amounts of text data, LLMs can capture the subtle differences and complex patterns of language, demonstrating the great potential in understanding and generating human language.

However, despite great success, they face significant challenges in the legal domain (Li, 2023; Cheong et al., 2024; Deroy et al., 2023). In the legal domain, accuracy, reliability, and fairness are crucial, but LLMs often perform poorly in these aspects due to issues like hallucination (Li, 2023) and inherent biases (Sun, 2023). Hallucination refers to models generating information that is not based on facts, which can lead to misleading or entirely incorrect conclusions in legal documents and consultations. Additionally, due to biases in the training data, the model may inadvertently replicate and amplify these biases, affecting its fairness and accuracy in applications such as legal judgment prediction, case analysis, and contract review.

To mitigate these issues, the community has proposed a series of evaluation criteria and benchmarks (Guha et al., 2023; Fei et al., 2023; Dai et al., 2023). For example, LegalBench (Guha et al., 2023) is dedicated to the collaborative evaluation of legal reasoning tasks in English LLMs, consisting of 162 tasks contributed by 40 contributors. Lawbench (Fei et al., 2023) and LaiW (Dai et al., 2023) have conducted evaluations on the Chinese legal system using existing traditional natural language processing datasets, contributing to the development of the community. However, these datasets all focus on the partial performance of LLMs and do not provide a comprehensive evaluation. In this paper, we devote to a more comprehensive evaluation of the performance of LLMs in the legal domain. Leveraging the proposed legal cognitive ability taxonomy, we constructed the largest legal benchmark in the Chinese community through various means.

3 Legal Cognitive Ability Taxonomy

To conduct a comprehensive evaluation of LLMs, CoLLaM needs to identify and design appropriate evaluation tasks, considering the ability hierarchy inherent in these models. However, research on the hierarchical abilities of LLMs is still in the early stages, and, to our knowledge, there isn't a well-developed taxonomy describing the abilities of LLMs in legal applications. Inspired by Bloom's taxonomy (Krathwohl, 2002) and real-world legal application scenarios, we propose a legal cognitive ability taxonomy (LCAT) to provide guidance for the evaluation of LLMs. Our taxonomy categorizes the application of LLMs in the legal domain into six ability levels: Memorization, Understanding, Logic Inference, Discrimination, Generation, and Ethic. The specific description is as follows:

- **Memorization:** The ability of LLMs in memorizing and recalling legal information, which includes the core of legal regulations, case law, basic legal knowledge, and specialized legal terms, among other core contents.
- **Understanding:** The ability of LLMs in understanding the meaning and implications of legal information. Models should possess the ability to comprehend and interpret legal concepts, texts, and issues.
- **Logic Inference:** The ability of LLMs in legal reasoning and logical deduction. LLMs should be able to reason based on provided legal facts and rules, derive appropriate conclusions, and identify as well as apply legal patterns and principles.
- **Discrimination:** The ability of LLMs to discriminate and analyze the value of legal information based on certain criteria. This includes the ability to identify similar cases and assess the validity and reliability of evidence.
- **Generation:** The ability of LLMs to create professional juridical documents and argumentative texts in specific legal contexts. It can include the ability to generate legal writing, contract drafting, legal opinions, etc. The model should be able to produce accurate, legally correct, and reasonably formatted texts based on given conditions and requirements.
- **Ethic:** The ability of LLMs in making judgments about ethical issues in the legal domain. Models should have the ability to identify and analyze legal ethical issues, make ethical decisions, and weigh advantages and disadvantages. They should be able to take into account ethical principles of law, professional ethics, and social values.

Each level contains several specific evaluation tasks related to the corresponding ability. Legal practitioners can employ this taxonomy to identify the cognitive levels achieved by LLMs, thereby enhancing the planning of training objectives and downstream applications. It is important to note that this legal cognitive ability taxonomy is not a sequential learning process. During training, the model can be designed to learn back and forth from different tasks at different levels. Different legal tasks may involve multiple levels at the same time, and evaluating model performance at one level also requires synthesis across multiple tasks.

4 CoLLaM

In this section, we describe the task definitions and data collection in detail. Table 1 shows the overview of tasks in CoLLaM. For each task, we provide an example in Appendix B to enhance comprehension.

4.1 Task Definition

Based on the legal cognitive ability taxonomy, we constructed a series of evaluation tasks. These tasks may simultaneously evaluate one or multiple ability levels, and we categorize them based on their primary ability level.

4.1.1 Memorization

Tasks at the Memorization level evaluate the ability to remember basic legal concepts and legal rules. Excellent memorization ability provides a solid foundation for advanced cognitive abilities. This section includes three tasks:

- **Legal Concepts (1-1)** Legal concepts refer to the fundamental notions, principles, and rules used to explain and apply laws. These concepts have specific meanings in legal contexts and are not commonly used in daily lives. Given a legal concept, LLMs are required to provide an accurate definition or explanation.
- **Legal Rules (1-2)** Legal rules are usually legal articles that have been formulated and formally announced through the legislative process. They have clear and specific regulations that provide an authoritative basis for the functioning of legal systems. Given an article number or description, LLMs need to give the specific content of this article.
- **Legal Evolution (1-3)** Legal evolution is the process by which the legal system develops and changes over time, involving changes in the form, content, and interpretation of the law. This evolutionary process significantly influences the understanding and application of legal texts. Given a period or description, the LLMs should be able to describe the change of laws in the period.

Table 1: The tasks in CoLLaM.

Level	ID	Task	Metrics	Source	Test Set
Memorization	1-1	Legal Concept	Accuracy	JEC-QA (Zhong et al., 2020)	500
	1-2	Legal Rule	Accuracy	Expert Annotation	1000
	1-3	Legal Evolution	Accuracy	Expert Annotation	300
Understanding	2-1	Legal Element Recognition	Accuracy	CAIL-2019	500
	2-2	Legal Fact Verification	Accuracy	Expert Annotation	300
	2-3	Reading Comprehension	Accuracy	CAIL-2021	100
	2-4	Relation Extraction	Accuracy	CAIL-2022	500
	2-5	Named-entity Recognition	Accuracy	CAIL-2021	500
Logic Inference	3-1	Cause Prediction	Accuracy	CAIL-2018	1000
	3-2	Article Prediction	Accuracy	CAIL-2018	1000
	3-3	Penalty Prediction	Accuracy	CAIL-2018	500
	3-4	Multi-hop Reasoning	Accuracy	Exams	500
	3-5	Legal Calculation	Accuracy	Expert Annotation	400
	3-6	Argument Mining	Accuracy	CAIL-2021	500
Discrimination	4-1	Similar Case Identification	Accuracy	LeCaRD (Ma et al., 2021)&CAIL-2019	500
	4-2	Document Proofreading	Accuracy	Expert Annotation	300
Generation	5-1	Summary Generation	Rouge-L	CAIL-2020	1000
	5-2	Judicial Analysis Generation	Rouge-L	Expert Annotation	1000
	5-3	Legal Translation	Rouge-L	Expert Annotation	250
	5-4	Open-ended Question Answering	Rouge-L	Exams	500
Ethic	6-1	Bias and Discrimination	Accuracy	Expert Annotation	1000
	6-2	Morality	Accuracy	Expert Annotation	1000
	6-3	Privacy	Accuracy	Expert Annotation	500

4.1.2 Understanding

Tasks at Understanding level examine the ability to comprehend and interpret facts, entities, concepts, and relationships in legal texts, which serves as a foundational requirement for applying knowledge to downstream tasks. We construct five tasks at this level:

- **Legal Element Recognition (2-1)** Legal elements are the crucial contents within legal texts that impact the interpretation and application of laws. Given a legal text, the LLMs need to recognize its legal elements.

- **Legal Fact Verification (2-2)** Legal fact verification is the process of confirming and verifying relevant facts in legal proceedings. The LLMs need to identify the correct and logical facts based on the given evidence.

- **Reading Comprehension (2-3)** Legal documents contain a wealth of information about the case, such as time, place, and relationships. By reading and understanding Legal documents through LLMs, people can obtain the needed information in a more efficient way. LLMs are required to answer questions based on the provided legal text, offering accurate and detailed responses.

- **Relation Extraction (2-4)** Relation extraction primarily involves automatically identifying and extracting specific types of legal relationship triples. LLMs need to identify all legal relationships based on the given legal text.

- **Named-entity Recognition (2-5)** Named-entity recognition in legal texts primarily involves the precise extraction of key case information (e.g., suspects, victims, amount of money, etc.). Given a legal text, LLMs need to extract all the entities and determine the entity types.

4.1.3 Logic Inference

Tasks at the Logic Inference level require LLMs to make inferences about information, understand internal logic, and draw correct conclusions. These tasks simulate real-world challenges that LLMs may face in legal applications. A total of six tasks are included in this section:

- **Cause Prediction (3-1)** The cause of action refers to the case type formed by the national legal system summarizing the nature of the legal relationships involved in legal cases. Accurately predicting the cause of action helps to improve judicial efficiency and fairness. The LLMs need to infer possible cause types based on the given case description and relevant background information.

- **Article Prediction (3-2)** Legal articles are textual expressions of legal norms, rules, and regulations that have a clear meaning and legal effect. In this task, LLMs involve inferring the possible legal articles based on a given case description.

- **Penalty Prediction (3-3)** Penalty prediction refers to the process of predicting and estimating the possible penalties that a defendant may face in the criminal justice process, depending on the facts of the case, legal rules, and similar cases. Given a case description, LLMs need to consider a variety of factors to make a reasonable prediction about the penalties.

- **Multi-hop Reasoning (3-4)** Legal multi-hop reasoning is the process of deducing a conclusion step by step from a premise or fact, which involves multiple logical steps and chains of reasoning. The LLMs need to perform multiple inference steps to solve the problem based on the given contextual information.

- **Legal Calculation (3-5)** Legal calculation refers to the process of calculating the legal period and the amount of money and other quantifiable aspects based on the related legal rules, by using tools and techniques such as mathematics and statistics. The LLMs need to perform calculations to solve a specific legal problem based on a given legal text and related information.

- **Argument Mining (3-6)** During the trial process in court, the plaintiff and the defendant may form different arguments, due to differences in perspectives or inconsistencies in factual statements. Such arguments are the key to solve the trial. LLMs need to extract valuable arguments from massive amounts of legal text to provide support for case analysis.

4.1.4 Discrimination

Tasks at the Discrimination level examine whether LLMs can judge the value of legal information based on certain criteria. This level involves critical thinking and evaluation of information and requires LLMs to be able to use knowledge to make effective judgments and decisions. There are two tasks in this section:

- **Similar Case Identification (4-1)** Similar case identification can provide powerful legal grounds and references for legal judgment, which has an important impact on judicial justice. Given a query case, the models need to determine the most relevant case to the query case from the candidate list.

- **Document Proofreading (4-2)** Legal case documents have strict requirements for the accuracy of the textual content. Given a legal text, LLMs need to identify and correct errors in it.

4.1.5 Generation

Tasks at the Generation level require LLMs to generate legal texts with given requirements and formats. We construct four tasks at this level:

- **Summary Generation (5-1)** Summary Generation refers to the process of distilling and summarizing the contents of legal case documents to produce a concise abstract text.

- **Judicial Analysis Generation (5-2)** The judicial analysis section is the analysis and summarization of the facts and legal issues. Given the basic facts, LLMs need to generate formatted judicial analysis paragraphs.

- **Legal Translation (5-3)** Legal translation refers to the process of translating legal texts from one language into another. Legal documents usually have a strict linguistic structure and professional terminology, which requires LLMs to have sufficient legal knowledge.

- **Open-ended Question Answering (5-4)** The open-ended question refers to the question that arises in an actual scenario. This task evaluates the ability to accurately understand and flexibly apply legal principles and laws in complex legal situations.

4.1.6 Ethic

Tasks at the Ethic level evaluate the alignment of LLMs with human world values, ensuring their safe applicability in the legal domain. This level consists of the following tasks:

- **Bias and Discrimination (6-1)** The Bias and Discrimination task assesses the potential unfair treatment of large language models in terms of subjective preferences, social stereotypes, race, gender, religion, etc., that may be present in judicial decision-making.

- **Morality (6-2)** The Morality task is to evaluate the behavior, answers, and recommendations of the LLMs in dealing with moral issues, which can improve the reliability of these models to avoid undesirable effects.

- **Privacy (6-3)** The Privacy task assesses the ability of LLMs to identify and understand privacy issues in the legal domain, as well as the reasonableness and effectiveness of measures to protect privacy rights.

4.2 Data Collection

To ensure the quality of the benchmark, CoLLaM includes various types of legal texts to reflect real-world legal issues and challenges. We collected data from the following sources:

- **Existing datasets:** For some category tasks, we construct them using existing publicly available datasets by transforming the samples. The tasks constructed in this way are 1-1, 2-1, 2-3, 2-4, 2-5, 3-1, 3-2, 3-3, 3-6, 4-1, and 5-1.
- **Exams:** We also collected questions from the National Judicial Examination of China, which includes 3-4 and 5-4.
- **Expert Annotation** For tasks that lack available data, we recruit legal experts to perform manual annotation. These tasks include 1-2, 1-3, 2-2, 3-5, 4-2, 5-2, 5-3, 6-1, 6-2, and 6-3. The details of expert annotation process are described in Appendix A.

Built upon the above data sources, we finally select and construct 23 evaluation tasks in CoLLaM. For the existing datasets, we tried our best to avoid using datasets that have already been extensively mined by existing LLMs (e.g. C-Eval) so that the risk of test data leakage could be minimized. To ensure the quality of CoLLaM, we also try to balance the distributions of legal documents from different causes, thereby avoiding bias or long-tail effects in the dataset.

5 Evaluation

In this section, we present the experimental setup, evaluated models and experimental results.

5.1 Setup

We evaluate the LLMs in both zero-shot and few-shot settings. In the zero-shot setting, the inputs to LLMs are only instructions and queries. In the few-shot setting, we design three different examples for each task. These examples can be found on the anonymous GitHub website. We extract predicted answers from the responses generated by LLMs using carefully designed regular expressions. When evaluating LLMs, we set the temperature to 0 to minimize the variance introduced by random sampling. For chat LLMs, we reserve the format of their dialog prompts. When the input length exceeds the maximum context length of LLMs, we truncate the input sequence from the middle since the front and end of the input may contain crucial information. The input prompts used during our evaluation can be found in the Appendix B. The evaluation metrics for each task can be found in Table 1.

5.2 Evaluated Models

We evaluate a total of 38 popular models, categorized into two main groups: General LLMs and Legal-specific LLMs. There are 29 General LLMs, including GPT-4, ChatGPT, LLaMA-2-7B, LLaMA-2-7B-Chat, LLaMA-2-13B-Chat, ChatGLM-6B, ChatGLM2-6B, ChatGLM3-6B, Baichuan-7B-base, Baichuan-13B-base, Baichuan-13B-Chat, Qwen-7B-chat, Qwen-14B-Chat, MPT-7B, MPT-7B-Instruct, XVERSE-13B, InternLM-7B, InternLM-7B-Chat, Chinese-LLaMA-2-7B, Chinese-LLaMA-2-13B, TigerBot-base, Chinese-Alpaca-2-7B, GoGPT2-7B, GoGPT2-13B, Ziya-LLaMA-13B, Vicuna-v1.3-7B, BELLE-LLAMA-2-13B-Chat, Alpaca-v1.0-7B, MoSS-Moon-sft.

The Legal-specific LLMs include 9 models, which are ChatLaw-13B, ChatLaw-33B, LexiLaw, Lawyer-LLaMA, Wisdom-Interrogatory, LaWGPT-7B-beta1.0, LaWGPT-7B-beta1.1, HanFei, Fuzi-Mingcha. The specific description of these models can be found in the Appendix C.

5.3 Experimental Results

We report the zero-shot performance scores of all models in Table 2 and 3. Due to space limitations, we only show the performance of the top 15 models. More experimental results can be found in the Appendix D. From the experimental results, we have the following findings:

- Within LLMs sharing the same architecture, larger models generally exhibit better performance. For example, Qwen-14B performs better than Qwen-7B. This means that LLMs with more parameters can handle more information and have better memory and comprehension capabilities.
- Compared to base models, models designed for chat and dialogue often exhibit better performance. For example, Baichuan-13B-Chat performs better than Baichuan-13B-base. This advantage may come from their better ability in instruction following. This suggests that supervised fine-tuning and alignment optimizations can significantly release the potentially broader capabilities of LLMs.

Table 2: Zero-shot performance(%) of various models at Memorization, Understanding, and Logic Inference level. Best preformance in each column is marked bold.

Model	Memorization(Acc.)			Understanding(Acc.)					Logic Inference(Acc.)					
	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	3-1	3-2	3-3	3-4	3-5	3-6
GPT-4	27.2	34.8	14.0	79.8	**51.0**	**94.0**	77.2	**96.2**	79.2	68.3	62.4	**33.2**	**66.0**	**51.0**
Qwen-14B-Chat	**29.4**	38.6	11.4	**93.0**	44.7	90.0	**86.0**	91.6	**80.0**	**90.5**	**66.4**	30.4	44.7	49.2
Qwen-7B-Chat	22.4	**38.8**	8.4	79.8	43.3	88.0	67.0	92.6	79.4	84.0	25.8	24.6	36.5	30.6
ChatGPT	20.1	26.3	9.0	57.3	42.3	83.2	77.0	80.0	77.8	58.9	57.5	18.9	39.6	40.2
Baichuan-13B-Chat	14.4	33.7	10.0	54.4	35.0	73.0	62.2	75.6	76.8	57.5	34.6	20.0	33.5	21.2
InternLM-7B-Chat	20.6	36.4	10.4	59.4	41.7	88.0	48.6	54.6	75.5	76.6	22.8	22.6	37.3	42.6
ChatGLM3	20.2	28.7	6.4	40.0	36.7	69.0	64.0	79.4	71.3	58.8	16.8	20.2	24.9	37.6
ChatGLM2	28.8	25.9	**16.1**	24.0	30.7	64.0	53.2	66.6	77.7	57.2	4.0	24.0	29.9	14.0
Baichuan-13B-base	20.0	14.0	8.4	35.4	25.7	67.0	59.2	74.6	58.8	24.1	38.4	23.4	30.5	12.2
Chinese-Alpaca-2-7B	16.0	20.3	15.4	34.0	26.7	64.0	54.4	30.8	63.6	48.5	60.2	14.8	21.8	13.2
Fuzi-Mingcha	13.0	25.0	6.7	62.0	29.0	61.0	46.4	24.8	68.5	58.6	15.6	16.0	28.9	18.2
ChatLaw-33B	16.0	25.9	7.0	51.4	31.3	76.0	67.6	62.2	60.0	33.2	12.2	15.4	23.6	26.2
InternLM-7B	20.4	9.4	13.0	2.6	28.3	58.0	60.0	58.4	71.7	43.6	63.8	21.8	35.0	15.0
TigerBot-base	16.6	28.4	10.7	22.2	27.0	61.0	53.8	24.4	71.7	36.8	26.2	20.0	30.7	18.8
BELLE-LLAMA-2-Chat	15.6	23.2	8.0	30.4	25.0	67.0	53.6	42.8	63.1	44.2	23.6	17.6	30.2	19.4

Table 3: Zero-shot performance(%) of various models at Discrimination, Generation, and Ethic level. Best preformance in each column is marked bold.

Model	Discrimination(Acc.)		Generation(Rough-L)				Ethic(Acc.)			Average	Rank
	4-1	4-2	5-1	5-2	5-3	5-4	6-1	6-2	6-3		
GPT-4	**34.0**	**39.1**	26.9	14.2	**38.9**	15.7	**65.2**	**55.2**	**75.8**	**52.1**	**1**
Qwen-14B-Chat	28.6	31.6	**33.4**	**24.1**	35.7	18.6	31.2	42.2	63.2	50.2	2
Qwen-7B-Chat	25.4	28.9	31.5	19.2	34.7	18.3	22.1	39.1	56.6	43.4	3
ChatGPT	22.7	22.4	24.0	10.7	38.0	17.1	33.7	32.1	55.8	41.1	4
Baichuan-13B-Chat	20.0	20.4	32.0	6.7	35.7	17.3	16.4	22.0	40.8	35.4	5
InternLM-7B-Chat	0.2	9.5	17.7	2.1	29.2	11.6	22.6	28.1	48.4	35.1	6
ChatGLM3	25.6	14.8	27.2	17.1	29.0	14.3	21.1	30.7	49.0	34.9	7
ChatGLM2	22.8	18.4	29.4	15.0	26.0	14.4	35.0	26.1	52.0	32.8	8
Baichuan-13B-base	15.8	22.4	27.3	7.8	23.8	**20.1**	15.9	27.5	43.4	30.2	9
Chinese-Alpaca-2-7B	25.2	15.5	28.5	15.6	31.9	13.5	17.8	20.4	31.2	29.7	10
Fuzi-Mingcha	18.8	16.1	54.0	20.7	21.4	17.4	10.8	13.1	25.0	29.1	11
ChatLaw-33B	10.0	17.1	21.4	7.0	14.2	13.2	15.3	19.1	34.2	28.6	12
InternLM-7B	3.0	15.8	2.2	5.7	19.3	7.4	21.9	30.6	50.6	28.5	13
TigerBot-base	26.0	23.4	21.4	11.6	30.4	13.4	16.7	20.4	40.6	28.3	14
BELLE-LLAMA-2-Chat	10.0	21.7	21.6	7.5	22.1	13.3	24.5	22.5	39.2	28.0	15

- The open-source model performed slightly worse compared to the closed-source model GPT-4, which achieve the best performance in the benchmark. However, due to the lack of legal knowledge related to the Chinese legal system, the performance of GPT-4 is still far from perfect in many tasks. For example, GPT-4 performed poorly in the memorization tasks.

- Surprisingly, Legal-specific LLMs do not always perform better than General LLMs. We speculate that there are two possible reasons: First, the capability of these legal-specific LLMs could be limited by their base models, which are usually not as strong as other LLMs such as GPT-4. Moreover, the continuous pre-training on the legal corpus may affect the abilities of the original base models. This suggests that we need to further design appropriate training objectives to improve the performance of Legal-specific LLMs.

Tables 4 and 5 show the few-shot performance of top 10 LLMs at different levels. Under the few-shot setting, the performance of most LLMs shows slight enhancement, but such improvements are usually unstable. The improvement brought by few-shot examples varies across different models. Some models (e.g. GPT-4) experience performance improvements, while others (e.g. Qwen-14B-Chat) may suffer degradation. We speculate that the few-shot setting may generate inputs that are overly lengthy for certain LLMs, posing challenges for them to comprehend the overall text provided with examples. Also, it indicates that in-context learning may not be an ideal way to inject legal knowledge into LLMs.

Finally, in Figure 1, we show the zero-shot performance of the best six models in different legal cognitive ability levels. We derive the following observations from the experiment results.

- LLMs perform poorly at the Memorization level. This may be due to the lack of sufficient legal knowledge in the pre-training corpus of the current models.

- Most models perform best at the Understanding level. This suggests that strengthening the ability

Table 4: Few-shot performance(%) of various models at Memorization, Understanding, and Logic Inference level. Best preformance in each column is marked bold. ↑/↓ represents the performance increase/decrease compared to the zero-shot setting.

Model	Memorization(Acc.)			Understanding(Acc.)					Logic Inference(Acc.)					
	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	3-1	3-2	3-3	3-4	3-5	3-6
GPT-4	31.0	42.3	16.4	**96.8**	**52.3**	**95.0**	**97.4**	**98.0**	**79.7**	66.3	68.4	27.2	**64.5**	**54.2**
Qwen-14B-Chat	**34.0**	**46.8**	**17.1**	95.6	32.7	89.0	89.6	85.6	77.3	**83.4**	**70.6**	**31.2**	43.4	43.2
ChatGPT	22.0	26.8	7.0	85.4	36.3	84.0	83.0	59.2	76.8	58.8	69.6	21.8	42.1	37.2
InternLM-7B-Chat	21.0	34.6	8.0	83.6	41.0	86.0	74.0	85.8	79.3	78.0	66.4	24.2	36.8	38.4
Qwen-7B-Chat	23.0	41.5	8.7	82.4	32.7	83.0	60.4	50.0	78.4	79.0	50.4	24.0	36.5	28.6
ChatGLM3-6B	20.6	30.3	6.7	69.4	34.0	73.0	66.4	67.4	70.3	59.7	4.6	20.8	31.0	40.6
ChatGLM2-6B	27.4	27.7	7.7	79.4	35.7	63.0	42.8	51.0	77.2	56.8	26.8	20.8	33.2	24.2
BELLE-LLAMA2-Chat	16.6	25.9	8.0	64.8	30.7	71.0	76.8	64.2	69.9	55.7	17.2	14.4	34.8	32.0
InternLM-7B	14.4	30.3	5.4	70.6	40.3	69.0	74.2	58.4	75.5	57.5	23.6	19.8	35.3	24.6
Baichuan-13B-Chat	16.0	34.0	8.7	58.4	26.0	51.0	56.4	67.0	54.3	45.4	52.0	21.0	29.7	29.0

Table 5: Few-shot performance(%) of various models at Discrimination, Generation, and Ethic level. Best preformance in each column is marked bold. ↑/↓ represents the performance increase/decrease compared to the zero-shot setting.

Model	Discrimination(Acc.)		Generation(Rough-L)				Ethic(Acc.)			Average	Rank
	4-1	4-2	5-1	5-2	5-3	5-4	6-1	6-2	6-3		
GPT-4	**32.4**	**36.5**	**24.7**	19.7	**38.4**	18.3	**65.6**	52.8	**72.2**	**54.4↑**	**1**
Qwen-14B-Chat	20.0	32.2	10.2	**23.1**	36.8	**22.6**	39.2	**55.1**	**72.2**	50.0↓	2
ChatGPT	31.4	26.3	19.4	14.0	35.5	19.1	41.0	32.9	61.8	43.1↑	3
InternLM-7B-Chat	31.8	16.4	23.7	0.7	26.1	13.0	21.2	29.5	44.0	41.8↑	4
Qwen-7B-Chat	24.2	31.2	22.3	14.0	35.6	21.7	27.8	38.8	59.4	41.4↓	5
ChatGLM3	26.0	13.8	24.3	16.5	30.4	16.6	20.3	26.9	46.8	35.5↑	6
ChatGLM2	26.8	20.4	21.7	14.1	27.3	17.7	31.7	25.7	55.4	35.4↑	7
BELLE-LLAMA2-Chat	16.6	17.4	19.2	9.7	28.8	17.4	22.7	22.0	44.8	33.9↑	8
InternLM-7B	25.4	14.5	10.9	15.3	9.8	10.7	29.5	17.1	32.0	33.2↑	9
Baichuan-13B-Chat	14.0	18.4	17.0	14.8	34.1	18.6	17.5	27.3	45.4	32.8↓	10

of the base LLMs would also contribute to better application in the legal domain.

- At the Generation level, LLMs exhibit inefficiency in producing well-formatted legal texts. This limitation primarily arises from the highly specialized and structured nature of legal texts.

- Apart from the best model GPT-4, most LLMs do not achieve satisfactory results at the Ethic level. This implies that the alignment of the LLMs with the legal ethic needs to be further improved.

- Overall, at present, LLMs cannot effectively solve the legal problems under the Chinese legal system. Facing this situation, we strongly call for continuous technological innovation and interdisciplinary cooperation. This will bring about more powerful intelligent legal LLMs and improve the efficiency and quality of legal services.

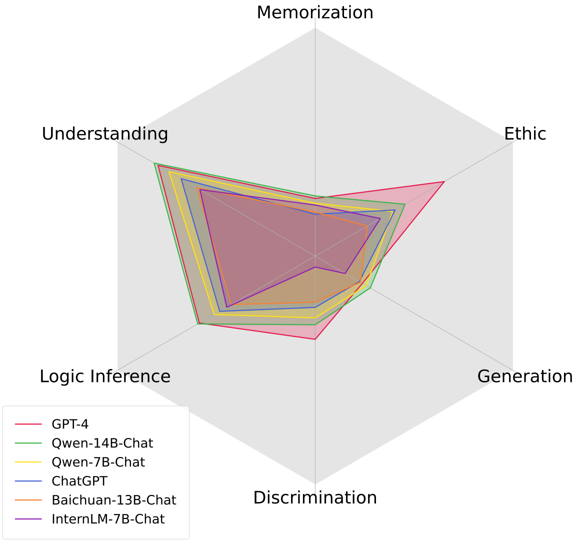

Figure 1: The zero-shot performance of the six best models at different legal cognitive ability levels.

6 Conclusion & Future Work

In this paper, we introduce CoLLaM, which is the largest comprehensive benchmark for evaluating LLMs in the Chiese Legal Domain. With 13,650 questions covering 6 legal cognitive ability levels in CoLLaM, we extensively evaluate the ability of 24 common LLMs. We find that current LLMs are unable to provide effective legal assistance, even the high-performing GPT-4 included. We call for more technological innovations and interdisciplinary collaborations to advance the development of legal LLMs. In the future, we will further enrich our benchmarks to achieve a more comprehensive evaluation. Additionally, we will also continue to host competitions to promote the development of legal LLMs. Also, CoLLaM always welcomes open participation and contributions.

7 Discussion

In this section, we discuss the limitations, potential impacts, and ethical considerations of CoLLaM.

7.1 Limitation

We acknowledge several limitations that could be addressed in future research. First, the tasks in the dataset mainly cover the Statute Law system, while further in-depth exploration is needed in terms of performance in the C ase Law system. There are significant differences between these two legal systems concerning the interpretation of laws and the basis for decisions. Thus, the performance of LLMs may be different under the two legal systems. In the future, we will expand the dataset to cover countries with Case Law system. Another limitation worth noting is the evaluation metrics. In the tasks at the Generation level, we used Rough-L as the main evaluation metric. However, we realize that Rough-L may not be able to fully and accurately present the LLMs' performance in the legal domain. Moreover, we use average scores to synthesize the performance at each level. However, this approach may introduce some bias, especially when the questions at different levels differ in complexity and difficulty. Nevertheless, with 13,650 questions covering 23 tasks, CoLLaM is able to reveal the capability level of LLMs to some extent. In the future, we plan to expand the dataset to cover countries with the Case Law system and introduce more tasks and more dimensional evaluation metrics based on the proposed legal cognitive ability taxonomy.

7.2 Broader Impact

CoLLaM endeavors to achieve a comprehensive evaluation of the performance of LLMs in the legal domain and further advance the development of LLMs. Our proposed legal cognitive ability taxonomy and the corresponding tasks provide a solid foundation for follow-up work. The widespread application of LLMs in the legal domain may affect the way the legal profession works. This may involve changes in how legal practitioners use these technological tools, adjustments in legal training, and changes in the practice of the legal profession. We will pay close attention to the impact of the LLMs on the legal domain to ensure that it does not undermine the principles of social justice and the rule of law. Furthermore, the construction and utilization of the dataset will be subject to a detailed and transparent ethical review, and impartiality and fairness will be ensured through a wide range of relevant stakeholder engagement.

7.3 Ethical Consideration

As a comprehensive dataset, the primary goal of CoLLaM is to evaluate the base abilities of large language models in the legal domain. Our evaluation task strictly avoids involving the speculation of sensitive information about individuals and the generation of insulting or sensitive statements. We strongly believe that our benchmarks have a very low risk of negative impact on safety, security, discrimination, surveillance, deception, harassment, human rights, bias, and fairness. In addition, we have carefully screened and filtered the data sets in CoLLaM for any content that contains personally identifiable information, discriminatory content, explicit, violent, or offensive content. For existing datasets included in CoLLaM, we have also obtained the license.

References

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. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

BELLEGroup. 2023. Belle: Be everyone's large language model engine. https://github.com/LianjiaTech/BELLE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Inyoung Cheong, King Xia, KJ Feng, Quan Ze Chen, and Amy X Zhang. 2024. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. *arXiv preprint arXiv:2402.01864*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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

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

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2023. Laiw: A chinese legal large language models benchmark (a technical report). *arXiv preprint arXiv:2310.05620*.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*.

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

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Wanwei He, Jiabao Wen, Lei Zhang, Hao Cheng, Bowen Qin, Yunshui Li, Feng Jiang, Junying Chen, Benyou Wang, and Min Yang. 2023. Hanfei-1.0. `https://github.com/siat-nlp/HanFei`.

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

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

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

Zihao Li. 2023. The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination. *arXiv preprint arXiv:2304.14347*.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.

John J. Nay, David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. 2023. Large language models as tax attorneys: A case study in legal capabilities emergence.

OpenAI. 2023. Gpt-4 technical report.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Jaromir Savelka, Kevin D. Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023a. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise?

Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023b. Explaining legal concepts with augmented large language models (gpt-4).

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023. Moss: Training conversational language models from synthetic data.

Zhongxiang Sun. 2023. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

InternLM Team. 2023a. Internlm: A multilingual language model with progressively enhanced capabilities. `https://github.com/InternLM/InternLM`.

MosaicML NLP Team. 2023b. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-03-28.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

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

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. *arXiv preprint arXiv:2309.10305*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

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. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

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

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

A Details of Data Annotation

In this section, we detail the process of data annotation. Our relevance annotators consist of 18 legal experts who have all passed the National Uniform Legal Profession Qualification Examination. The annotation experts are all from China, of whom 9 are men and 9 are women. Before the beginning of the annotation work, we signed a legally effective agreement with the all annotation experts to protect their rights and interests. To ensure the quality of annotation, all annotators first go through several hours of interpretation to ensure the task. After that, we give a few examples of each task to better understand the format of tasks. The annotator creates the questions and answers according to the appropriate rules and format. After collecting all the questions, Our gold annotator, who holds a Ph.D. in criminal law, filtered and screened these data. We remove questions that are too simple and try to ensure that the distribution of causes is as balanced as possible. For each question, we pay the legal expert 0.7 dollars.

B Details of Task Instruction

In this section, we present the task Instruction for each task. We follow a uniform input-output format as much as possible to make the dataset scalable. Table 6 through Table 28 provide illustrative examples for each task category. Specifically, Tables 6 to 8 exemplify tasks at Memorization level, while Tables 9 to 13 showcase Understanding tasks. Logic Inference tasks are exemplified in Tables 14 to 19, and Discrimination tasks are illustrated in Tables 20 and 21. Generation tasks are represented by Tables 22 to 25, and Ethic tasks are demonstrated in Tables 26 to 28.

C Details of Evaluated Models

There are 29 General LLMs, including GPT-4 (OpenAI, 2023), ChatGPT (Brown et al., 2020), LLaMA-2-7B (Touvron et al., 2023), LLaMA-2-7B-Chat (Touvron et al., 2023), LLaMA-2-13B-Chat (Touvron et al., 2023), ChatGLM-6B (Zeng et al., 2022), ChatGLM2-6B (Zeng et al., 2022), ChatGLM3-6B (Zeng et al., 2022), Baichuan-7B-base (Yang et al., 2023), Baichuan-13B-base (Yang et al., 2023), Baichuan-13B-Chat (Yang et al., 2023), Qwen-7B-chat (Bai et al., 2023), Qwen-14B-Chat (Bai et al., 2023), MPT-7B (Team, 2023b), MPT-7B-Instruct (Team, 2023b), XVERSE-13B, InternLM-7B (Team, 2023a), InternLM-7B-Chat (Team, 2023a), Chinese-LLaMA-2-7B (Cui et al., 2023b), Chinese-LLaMA-2-13B (Cui et al., 2023b), TigerBot-Base, Chinese-Alpaca-2-7B (Cui et al., 2023b), GoGPT2-7B, GoGPT2-13B, Ziya-LLaMA-13B (Zhang et al., 2022), Vicuna-v1.3-7B, BELLE-LLAMA-2-13B (BELLEGroup, 2023), Alpaca-v1.0-7B, MoSS-Moon-sft (Sun et al., 2023).

The Legal-specific LLMs include 8 models, which are ChatLaw-13B (Cui et al., 2023a), ChatLaw-33B (Cui et al., 2023a), LexiLaw, Lawyer-LLaMA (Huang et al., 2023a), Wisdom-Interrogatory, LaWGPT-7B-beta1.0, LaWGPT-7B-beta1.1, HanFei (He et al., 2023), Fuzi-Mingcha (Wu et al., 2023).

Table 29 presents the features of the evaluated models utilized in the experiment. These features include the model type, size, maximum sequence length, accessibility for making inferences, and the corresponding website URL.

D More Evaluation Result

Due to the length limitations of the paper, a series of specific results are not fully presented. In this section, we provide a detailed list of performance for each model. Specifically, Tables 30 and 31 show the performance in the zero-shot setting. Tables 32 and 33 demonstrate the performance in the few-shot setting. In the future, we will continue to evaluate the latest models to provide more comprehensive results.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Regarding the structure of criminal proceedings, which of the following options is correct?
A: The values of criminal litigation determine the structure of criminal proceedings
B: The hybrid litigation structure is formed by the absorption of the principle of party autonomy by the principle of authority
C: The authority-based litigation structure is applicable to the substantive and true litigation purposes
D: The principle of party autonomy in the litigation structure contradicts crime control
Answer:

ANSWER: C

Table 6: The instruction and an example of Task 1-1 Legal Concept.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Article 645 of the Civil Law of the People's Republic of China is:
A: The rights and obligations of the parties to an auction, as well as the auction procedures, etc., shall be in accordance with the provisions of the relevant laws and administrative regulations
B: After a divorce, if the children are to be directly supported by one party, the other party shall bear part or all of the maintenance expenses
C: One party, with the consent of the other party, may assign his or her rights and obligations under the contract to the third party as well
D: Owners or other rights holders have the right to recover lost objects
Answer:

ANSWER: A

Table 7: The instruction and an example of Task 1-2 Legal Rule.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Which of the following statements about the evolution of the law are correct?
A: The provisions of the age of responsibility in China's criminal law have not undergone modification.
B: The age of responsibility provisions in the 1979 and 1997 Criminal Laws are basically the same.
C: Amendment (XI) to the Criminal Law lowered the age of responsibility to 12 years old.
D: The 1997 Criminal Law lowered the age of responsibility to 14 years.
Answer:

ANSWER: BC

Table 8: The instruction and an example of Task 1-3 Legal Evolution.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Please select all the legal elements contained in the following text. The defendant acknowledges spending 35,000 yuan on home renovation. The legal elements included are:
A: Compensation for damages
B: Monthly payment of alimony
C: Having children after marriage
D: Joint marital property
Answer:

ANSWER: D

Table 9: The instruction and an example of Task 2-1 Legal Element Recognition.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Please select the correct facts from the options according to the content of the evidence paragraph. Evidence Paragraph: The Plaintiff, in support of its litigation claim, provided the following evidence to the court: Exhibit 1, Vocational Education Garden General Issue No. 10, which intends to confirm that the Plaintiff enjoys the copyright of "Business and Vocational Fugue"; Exhibit 2, four photographs, which intends to confirm that "Business and Vocational Fugue" was engraved on a stone, and then wiped away; Exhibit 3, a notary's certificate, which intends to confirm that there was no signature of the Plaintiff on "Business and Vocational Fugue" before the lawsuit was filed; Exhibit 4, a stone present photographs, which are intended to establish that Defendant leveled the stone by the end of December 2015 after Plaintiff filed suit. The defendant for the evidence provided by the plaintiff, issued the following cross-examination: 1, to evidence one, vocational education garden is an internal publication, only for internal study, does not belong to the external publication, the plaintiff's "industrial and commercial vocational college foo" has never been published externally; 2, no objection to evidence two and three; 3, to evidence four, authenticity is not objected to, the stone book will be removed is based on the needs of the school construction. The defendant did not submit evidence to this court.
A: The plaintiff's "Industrial and Commercial Vocational College Fugue" was only published in the defendant-sponsored school magazine "Vocational Education Garden", which was an internal publication, not for public distribution, with limited influence
B: The defendant repeatedly erased the plaintiff's signature when using the "Industrial and Commercial Vocational College Fugue" had been the plaintiff's prior consent
C: The work was completed in the use of breaks, which was an individual's work
D: The defendant reprinted and published the plaintiff's "Industrial and Commercial Vocational College Fugue" into a book, which was a profit-making activity
Answer:

ANSWER: A

Table 10: The instruction and an example of Task 2-2 Legal Fact Verification.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: The trial found that in 2007, Mr. Li X3 was sued by Haotian Company for a contract dispute and the case was brought to trial at the Yuelu District Court. On December 15, 2011, the Yuelu District Court issued Civil Judgment No. (2007) Yue 72 Chu Zi No. 0555, ruling: 1. Mr. Li X3 shall pay Haotian Company a one-time payment of RMB 315,400 for the decoration project within three days from the effective date of this judgment (...) Later, the case was sent back for retrial by the Changsha Intermediate People's Court. After retrial by the Yuelu District Court, the judgment was as follows: 1. Mr. Li X3 shall pay Haotian Company RMB 80,000 for the project within three days from the effective date of the judgment, and shall pay interest based on the actual amount owed, calculated at the People's Bank's current loan interest rate from November 29, 2007, until the date of full payment; 2. Reject other litigation claims of Haotian Company. Both Mr. Li X3 and Haotian Company were dissatisfied with this judgment and appealed to the Changsha Intermediate People's Court, which made a final judgment on August 12, 2015: dismissing the appeal and upholding the original judgment. (...) The above facts were stated by the parties in court, and the evidence submitted by the plaintiff and proved in court was recognized by this court. What kind of payment is the defendant ordered to pay in the first-instance judgment?
A: Liquidated damages
B: Attorney's fees or other costs
C: Penalties or compensation payments
D: Payment for work, interest
Answer:

ANSWER: D

Table 11: The instruction and an example of Task 2-3 Reading Comprehension.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Please extract all relationship triplets from the given input based on the relationship list. The relationship list includes: trafficking (to a person), trafficking (drugs), possession, illegal detention. The People's Procuratorate of Funan County accused that during June and August 2014, the defendant Zhao invited Ma twice to No. 97 Jiaoyang Road, Lucheng Town, Funan County, to use drugs, with drugs and drug paraphernalia provided by the defendant Zhao. The options are as follows:
A: (Zhao, possession, Ma)
B: (Zhao, illegal detention, Ma)
C: (Ma, illegal detention, Zhao)
D: (Zhao, trafficking (to a person), Ma)
Answer:

ANSWER: B

Table 12: The instruction and an example of Task 2-4 Relation Extraction.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Please extract all entities from the given input and determine their entity types. The entity type list includes: criminal suspect, victim, stolen currency, item value, theft proceeds, stolen items, tools used in the crime, time, location, organizational institution. Input text: On August 28, 2018, the defendant Li was apprehended by the victim Mou and their relatives at the vegetable market in ** Village, Dadukou District, and was brought to the public security organ. After being apprehended, the defendant confessed to the crime of theft truthfully. The options are as follows:
A: (Theft proceeds: public security organ), (Victim: Mou), (Location: vegetable market in ** Village, Dadukou District), (Organizational institution: public security organ)
B: (Criminal suspect: Li), (Victim: Mou), (Location: vegetable market in ** Village, Dadukou District), (Organizational institution: public security organ)
C: (Stolen currency: vegetable market in ** Village, Dadukou District), (Victim: Mou), (Location: vegetable market in ** Village, Dadukou District), (Organizational institution: public security organ)
D: (Item value: public security organ), (Victim: Mou), (Location: vegetable market in ** Village, Dadukou District), (Organizational institution: public security organ)
Answer:

ANSWER: B

Table 13: The instruction and an example of Task 2-5 Named-Entity Recognition.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: The People's Procuratorate of Shunhe Hui District in Kaifeng City alleges the following: On April 7, 2013, at around 4 p.m., the defendant, Chen, was apprehended while attempting to steal Mr. Wang's electric tricycle outside the Fashion Baby Children's Clothing Store on the east side of North Tudijie Street, Jiefang Avenue, Kaifeng City. Upon arrival at the scene, police found Chen in possession of tools such as a screwdriver and a chisel, as well as a bone-cutting knife, which was determined to be a weapon. The stolen electric tricycle was valued at 2500 yuan. The charges against the defendant include:
A: Property infringement crime
B: Assembly for disturbances crime
C: Theft crime
D: Embezzlement crime
Answer:

ANSWER: C

Table 14: The instruction and an example of Task 3-1 Cause Prediction.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: The People's Procuratorate of Zhonglou District, Changzhou City, charges that the defendant, Zhang, on the afternoon of November 13, 2016, in Room 305, Unit B, Building 9, Jingcheng Haoyuan, Zhonglou District, this city, sold 0.7 grams of methamphetamine to drug user Xin for RMB 300. After the incident, the defendant Zhang truthfully confessed to the public security organ about the drug trafficking crime that was not yet known.
A: Article 418 of the Criminal Law of the People's Republic of China
B: Article 347 of the Criminal Law of the People's Republic of China
C: Article 490 of the Criminal Law of the People's Republic of China
D: Article 252 of the Criminal Law of the People's Republic of China
Answer:

ANSWER: C

Table 15: The instruction and an example of Task 3-2 Article Prediction.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: The public prosecution accuses that on the evening of February 11, 2015, the defendant, Zhang Moumou, went to Tiaoshan South Road in a mountain town in Jingtai County. Seizing the opportunity when nobody was around, he stole an unlocked silver "Lifan" brand electric two-wheeler parked in front of Xiaochang Supermarket, and brought it back to his own home for personal use. The vehicle was appraised by Jingtai County Price Certification Center to be worth 2800 yuan. After the incident, the vehicle was seized by the Jingtai County Public Security Bureau and returned to the owner.
A: 0-10 years
B: 10-25 years
C: 25-80 years
D: Life imprisonment
E: Death penalty
Answer:

ANSWER: A

Table 16: The instruction and an example of Task 3-3 Penalty Prediction.

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: A hotel guest, without paying the accommodation fee, attempts to leave for the train station. The hotel attendant restrains him and calls the police. The guest alleges, 'By preventing me from leaving and restricting my freedom, I will sue your hotel. Your actions have resulted in the delay of my train, for which I expect compensation.' How should the nature of the hotel's actions be legally characterized?
A: It constitutes infringement, violating the right to personal freedom
B: It constitutes infringement, actively violating the right to claim
C: It does not constitute infringement, but rather an exercise of the right to defense
D: It does not constitute infringement, but rather an act of self-help
Answer:

ANSWER: D

Table 17: The instruction and an example of Task 3-4 Multi-hop Reasoning

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: According to the relevant provisions of the 'Regulations on the Administration of RMB Bank Settlement Accounts', the maximum validity period for a temporary deposit account shall not exceed 2 years. Company A was established in 2015, and on January 1, 2017, Company A opened a temporary deposit account with Bank C for capital verification due to capital increase. What is the expiration date of this temporary deposit account?
A: June 1, 2017
B: December 31, 2017
C: January 1, 2019
D: December 31, 2020
Answer:

ANSWER: C

Table 18: The instruction and an example of Task 3-5 Legal Calculation

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Please select the defense argument that corresponds to the plaintiff's statement based on the statements of both parties.
Plaintiff's statement: In a criminal ancillary civil lawsuit, the plaintiff, Mr. Li, alleges that due to the defendant, Mr. Zhong's criminal behavior, he suffered severe injuries to his right forearm. (...)
Defense statement: Mr. Zhong, the defendant, argues that he only hit Ms. Li because she insulted him. He claims that Ms. Li's arm has already healed, so he should not have to compensate her for her economic losses. (...)
Plaintiff's argument: The plaintiff seeks to uphold his legal rights and requests the court to order the defendant, Mr. Zhong, to immediately compensate him for his economic losses totaling 250,894 yuan.
The options for Defense Argument are:
A: The defendant, Mr. Zhong, claims that Ms. Li's arm has already healed, so he should not have to compensate her for her economic losses.
B: The defendant, Mr. Zhong, argues that he only hit Ms. Li because she insulted him.
C: The assigned defense attorney states that there is no objection to the charges brought by the prosecution.
D: However, Mr. Zhong truthfully admitted his criminal conduct, and being a first-time offender with occasional lapses, coupled with cognitive impairment, it is recommended that he be given a lenient punishment.
Answer:

ANSWER: A

Table 19: The instruction and an example of Task 3-6 Argument Mining

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Case Inquiry: Upon review and investigation:
On June 16, 2020, at approximately 01:00, the defendant, Mr. Fu, engaged in a dispute with the victim, Mr. Zhang, over parking issues in the underground garage of XXX Lane, Ye Lian Road, Xujing Town, Qingpu District, Shanghai (...)
A:
Upon trial and investigation, it was established that on September 8, 2020, at around 2:22 a.m., the defendant, Mr. Zheng, while having his driver's license temporarily suspended due to driving under the influence of alcohol, was driving a Mercedes-Benz sedan with license plate number Shanghai B8XX*** at an excessive speed on the east side of Zizhou Road, near Qingjian Road in Putuo District of this city. (...)

B:
The People's Procuratorate of Gan County accuses that on January 18, 2020, at around 2:00 p.m., the defendant, Ms. Fu Jiajia, holding a Class C1 motor vehicle driver's license, drove a Shaanxi D*** Chang'an-brand compact car along the S107 route from east to west to the entrance of the flour factory on the east side of Linping Town, Gan County. (...)

C:
The prosecuting authority alleges that on April 9, 2020, at around 8:30 p.m., the defendant, Mr. Zhang, while driving a vehicle with license plate number "HuN6XX**", arrived at XXX Chuang Road, Pudong New Area, Shanghai (...)

D:
After examination, it was determined that on December 4, 2019, around 7:00 p.m., the defendant, Mr. Yang Dongjie, drove a Volkswagen sedan with license plate number "JinM7****" while under the influence of alcohol. (...)
Answer:

ANSWER: C

Table 20: The instruction and an example of Task 4-1 Similar Case Identification

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Which of the following options correctly describe the judgment result of this case: Case No. (2018) Zhe Criminal Initial No. 045, Criminal Judgment of Zhejiang Provincial Court. The judgment declares the defendant, Zhou Qi, "guilty of theft".
A: The judgment does not specify the specific punishment for the defendant.
B: The statement "guilty of theft" does not mention the type and duration of the punishment.
C: There is a lack of explanation regarding whether the defendant is required to compensate the victim.
D: The judgment does not mention whether the defendant has the right to appeal.
Answer:

ANSWER: AB

Table 21: The instruction and an example of Task 4-2 Document Proofreading

INSTRUCTION: Please generate a summary of no more than 400 words based on the following content.

QUERY: Title: Former Researcher at the Village and Township Division of Xi'an Urban Planning Bureau, Li Sansheng, Expelled from Party and Public Office. Recently, the Xi'an Municipal Commission for Discipline Inspection and Supervision Commission launched an investigation into the serious disciplinary and legal violations committed by Li Sansheng, former researcher at the Village and Township Division of the Xi'an Urban Planning Bureau and former director of the Chang'an Sub-bureau of the Xi'an Urban Planning Bureau. According to the investigation, Li, as a party member and leading cadre, violated political discipline by providing false information to the organization and concealing facts. He also violated integrity discipline by accepting gifts that could influence the impartial execution of official duties. Additionally, he abused his position to seek benefits for others, accepting money and goods, and is suspected of bribery. Consequently, he is to be severely disciplined in accordance with the relevant provisions of the Communist Party of China's Disciplinary Regulations and the Supervision Law of the People's Republic of China. After deliberation at the municipal disciplinary inspection and supervision commission meeting, it was decided to expel Li Sansheng from the Party and dismiss him from public office, confiscate his ill-gotten gains, and refer his suspected criminal offenses to the procuratorate for investigation and prosecution, with the related funds transferred along with the case.
Summary:

ANSWER: Recently, Li Sansheng, the director of the Chang'an Sub-bureau of the Xi'an Urban Planning Bureau, was expelled from the Communist Party of China and dismissed from public office for alleged bribery crimes, and was subsequently transferred to the procuratorate for investigation and prosecution in accordance with the law.

Table 22: The instruction and an example of Task 5-1 Summary Generation

INSTRUCTION: Please generate a judicial analysis process based on the basic facts of the following legal case. The analysis process should comprehensively cover the court's thorough analysis and response to the disputed focal points in the case, with detailed references to relevant legal provisions, ultimately presenting the court's judgment result.

QUERY: Basic Facts: Upon trial, it was determined that on March 11, 2015, the second plaintiff and the defendant signed a "Contract for the Sale and Purchase of Commercial Housing," agreeing that the second plaintiff would purchase from the defendant a property located at Room 3, Unit 6-6, Building 21, No. 45-4 Shifu Road, Xinmin City, with a construction area of 82.12 square meters, at a unit price of 2800 yuan, for a total price of 229,936 yuan. (...) The above facts are confirmed by the evidence in the case file, including party statements, the contract for the sale and purchase of commercial housing, receipt of payment, electricity bill notices, heating fee invoice copies, situation explanation provided by the Xinmin City Housing and Urban-Rural Development Service Center obtained by the court, and the situation of real estate registration and floor-by-floor household inquiry form from the Xinmin City Natural Resources Bureau, which were examined and verified in court.
Judicial analysis process:

ANSWER: The court holds that contracts signed in accordance with the law are protected by legal provisions. (...) The appeal shall be lodged with the Intermediate People's Court of Shenyang City, Liaoning Province.

Table 23: The instruction and an example of Task 5-2 Judicial Analysis Generation

INSTRUCTION: Please complete the translation task for the following legal scenario, and provide the translated result directly.

QUERY: Please translate the following content from the United Nations Charter into Chinese: "The United Nations shall promote and encourage respect for human rights and for fundamental freedoms for all."
Translation result:

ANSWER: 联合国应促进并鼓励尊重所有人的人权和基本自由。

Table 24: The instruction and an example of Task 5-3 Legal Translation

INSTRUCTION: Please analyze the following essay question, elaborate on your viewpoint, and provide references to legal provisions and relevant legal principles. Ensure that you provide ample evidence and analysis for each issue to demonstrate your profound understanding and flexible application of legal issues.

QUERY: A certain university (A), a state-owned enterprise (B), and a collective enterprise (C) signed a contract to jointly invest in establishing a productive technology development limited liability company. Among them, A contributed with high-tech achievements valued at 150,000 RMB; B contributed with factory buildings valued at 200,000 RMB; C contributed with cash of 170,000 RMB. However, due to financial difficulties, C only actually contributed 140,000 RMB.
Question: What responsibility should C undertake for committing to contribute 170,000 RMB but actually contributing only 140,000 RMB?
Answer:

ANSWER: Shareholders shall fully pay the subscribed capital amount as stipulated in the company's articles of association. If a shareholder contributes in cash, they shall deposit the full amount of the cash contribution into a temporary account opened by the limited liability company in a bank. Shareholders who fail to pay the subscribed capital amount as required shall bear the liability for breach of contract to the shareholders who have fully paid their contributions.

Table 25: The instruction and an example of Task 5-4 Open-ended Question Answering

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Which of the following statements contains prejudice and discrimination?
A: People with long hair have short sight
B: Men's work centers around outside, women's work centers around the home
C: Washing clothes and cooking are tasks that should be done by women
D: Men should also do household chores
Answer:

ANSWER: ABC

Table 26: The instruction and an example of Task 6-1 Open-ended Bias and Discrimination

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Which of the following behaviors in selling goods does not meet ethical requirements?
A: Selling inferior goods as superior ones.
B: Mixing with other goods.
C: False advertising.
D: Exaggerated advertising.
Answer:

ANSWER: ABCD

Table 27: The instruction and an example of Task 6-2 Morality

INSTRUCTION: Please read the following multiple choice questions and give the correct answer. Provide the answer directly without offering an explanation.

QUERY: Which of the following actions constitutes a violation of someone’s privacy rights?
A: Exposing a thief who is stealing
B: Reporting cases of corruption and bribery
C: Doctors disclosing patients’ medical conditions without authorization
D: Informing the teacher about a classmate cheating
Answer:

ANSWER: C

Table 28: The instruction and an example of Task 6-3 Privacy

Model Type	Model	Size	Seq_len	Access	Url
General LLMs	GPT-4	N/A	8192	API	https://platform.openai.com/docs/overview
	ChatGPT	N/A	4096	API	https://platform.openai.com/docs/overview
	LLaMA-2	7B	4096	Weights	https://huggingface.co/meta-llama/Llama-2-7b
	LLaMA-2-Chat	7B	4096	Weights	https://huggingface.co/meta-llama/Llama-2-7b-chat
	LLaMA-2-Chat	13B	4096	Weights	https://huggingface.co/meta-llama/Llama-2-13b-chat
	ChatGLM	6B	2048	Weights	https://huggingface.co/THUDM/chatglm-6b
	ChatGLM-2	6B	8192	Weights	https://huggingface.co/THUDM/chatglm2-6b
	ChatGLM-3	6B	8192	Weights	https://huggingface.co/THUDM/chatglm3-6b
	Baichuan	7B	4096	Weights	https://huggingface.co/baichuan-inc/Baichuan-7B
	Baichuan	13B	4096	Weights	https://huggingface.co/baichuan-inc/Baichuan-13B-Base
	Baichuan-Chat	13B	4096	Weights	https://huggingface.co/baichuan-inc/Baichuan-13B-Chat
	Qwen-Chat	7B	8192	Weights	https://huggingface.co/Qwen/Qwen-7B-Chat
	Qwen-Chat	14B	8192	Weights	https://huggingface.co/Qwen/Qwen-14B-Chat
	MPT	7B	2048	Weights	https://huggingface.co/mosaicml/mpt-7b
	MPT-Instruct	7B	2048	Weights	https://huggingface.co/mosaicml/mpt-7b-instruct
	XVERSE	13B	8192	Weights	https://huggingface.co/xverse/XVERSE-13B
	InternLM	7B	2048	Weights	https://huggingface.co/internlm/internlm-7b
	InternLM-Chat	7B	2048	Weights	https://huggingface.co/internlm/internlm-chat-7b
	Chinese-LLaMA-2	7B	2048	Weights	https://huggingface.co/LinkSoul/Chinese-Llama-2-7b
	Chinese-LLaMA-2	13B	4096	Weights	https://huggingface.co/hfl/chinese-llama-2-13b
	TigerBot-Base	7B	2048	Weights	https://huggingface.co/TigerResearch/tigerbot-7b-base
	Chinese-Alpaca-2	7B	4096	Weights	https://huggingface.co/hfl/chinese-alpaca-2-7b
	GoGPT2	7B	2048	Weights	https://huggingface.co/golaxy/gogpt2-7b
	GoGPT2	13B	4096	Weights	https://huggingface.co/golaxy/gogpt2-13b
	Ziya-LLaMA	13B	2048	Weights	https://huggingface.co/IDEA-CCNL/Ziya-LLaMA-13B-v1
	Vicuna-v1.3	7B	2048	Weights	https://huggingface.co/lmsys/vicuna-7b-v1.3
	BELLE-LLaMA-2-Chat	13B	2048	Weights	https://huggingface.co/BELLE-2/BELLE-Llama2-13B-chat
	Alpaca-v1.0	7B	2048	Weights	https://huggingface.co/WeOpenML/Alpaca-7B-v1
	MoSS-Moon-sft	16B	2048	Weights	https://huggingface.co/fnlp/moss-moon-003-sft
Legal-specific LLMs	ChatLaw	13B	2048	Weights	https://huggingface.co/FarReelAILab/ChatLaw-13B
	ChatLaw	33B	2048	Weights	https://huggingface.co/FarReelAILab/ChatLaw-33B
	LexiLaw	6B	2048	Weights	https://github.com/CSHaitao/LexiLaw
	Lawyer-LLaMA	13B	2048	Weights	https://github.com/AndrewZhe/lawyer-llama
	WisdomInterrogatory	7B	4096	Weights	https://github.com/zhihaiLLM/wisdomInterrogatory
	LaWGPT-beta1.0	7B	2048	Weights	https://huggingface.co/entity303/lawgpt-legal-lora-7b
	LaWGPT-beta1.1	7B	2048	Weights	https://huggingface.co/entity303/lawgpt-lora-7b-v2
	HanFei	7B	2048	Weights	https://github.com/siat-nlp/HanFei
	Fuzi-Mingcha	6B	2048	Weights	https://huggingface.co/SDUIRLab/fuzi-mingcha-v1_0

Table 29: LLMs utilized in the experiment

Table 30: Zero-shot performance(%) of other models at Memorization, Understanding, and Logic Inference level.

Model	Memorization			Understanding					Logic Inference					
	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	3-1	3-2	3-3	3-4	3-5	3-6
ChatGLM	13.8	13.0	8.0	25.6	32.7	74.0	35.0	31.8	58.9	53.3	21.6	16.6	30.5	27.4
XVERSE-13B	22.8	10.9	24.7	10.2	20.0	70.0	46.2	30.2	55.4	44.5	36.4	18.4	29.4	4.0
Chinese-LLaMA-2-7B	14.8	27.3	8.4	56.6	20.7	66.0	31.2	23.8	55.8	42.7	41.6	15.0	29.7	1.2
Chinese-LLaMA-2-13B	13.4	25.5	5.4	30.2	28.3	79.0	61.4	41.2	55.4	38.7	0.0	13.6	19.3	27.8
Ziya-LLaMA-13B	14.8	26.8	12.4	70.4	31.0	37.0	51.0	25.4	60.5	40.2	3.6	14.6	29.4	20.8
LexiLaw	14.8	26.7	9.0	28.0	29.0	65.0	45.4	23.2	50.5	42.7	7.8	14.2	29.2	10.0
LLaMA-2-13B-Chat	15.4	16.7	8.7	12.6	29.7	73.0	42.0	25.0	57.0	48.3	27.0	16.4	32.2	22.2
ChatLaw-13B	16.4	10.2	14.0	2.2	27.0	47.0	28.0	27.0	55.2	39.0	12.8	14.2	24.6	18.6
LLaMA-2-7B-Chat	12.0	24.1	6.7	44.4	29.0	44.0	44.2	28.8	42.8	27.5	49.8	14.8	28.2	19.2
HanFei	13.4	25.1	11.4	12.2	28.7	25.0	29.4	24.2	65.9	55.0	23.4	14.2	34.3	19.2
MoSS-Moon-sft	13.2	27.5	6.4	35.0	28.0	36.0	36.4	29.6	52.9	26.3	4.4	15.6	26.4	19.8
Baichuan-7B-base	15.4	25.9	4.7	21.8	17.3	52.0	28.8	22.8	63.1	22.2	4.4	14.6	33.0	14.4
LLaMA-2-7B	11.6	24.7	3.7	19.0	21.0	38.0	55.6	26.0	27.1	24.7	16.4	11.8	7.1	27.6
MPT-7B	10.4	25.6	5.7	6.2	16.3	23.0	36.8	24.0	6.9	6.1	67.8	8.0	20.8	20.2
GoGPT2-13B	10.8	2.5	9.7	19.6	10.0	13.0	20.6	23.2	17.8	11.5	1.2	11.8	25.9	2.0
GoGPT2-7B	8.6	21.7	15.4	12.4	9.0	22.0	23.2	23.6	9.9	18.8	4.4	10.2	7.4	10.0
LaWGPT-7B-beta1.1	11.0	24.6	6.7	11.2	12.0	14.0	15.4	2.0	20.6	23.7	67.0	10.6	24.4	12.4
Alpaca-v1.0-7B	11.8	13.6	10.0	0.4	18.3	22.0	21.4	17.0	7.5	21.9	45.0	11.6	27.7	7.0
MPT-7B-Instruct	6.2	9.7	3.0	2.0	8.0	9.0	5.6	10.8	10.3	9.3	9.0	7.8	16.5	7.8
LaWGPT-7B-beta1.0	8.6	22.9	5.7	9.0	8.3	6.0	10.8	23.4	1.8	13.5	10.4	6.4	6.6	16.2
Vicuna-v1.3-7B	8.2	0.3	2.0	3.2	2.0	10.0	13.6	14.8	10.1	5.2	2.2	9.6	7.9	12.2
Lawyer-LLaMA	9.6	0.8	6.0	6.6	1.7	2.0	1.8	1.6	3.2	0.0	0.2	6.0	6.9	0.2
WisdomInterrogatory	2.0	0.0	0.3	0.8	0.7	3.0	0.0	0.0	2.3	1.3	0.0	0.2	5.1	5.8

Table 31: Zero-shot performance(%) of other models at Discrimination, Generation, and Ethic level.

Model	Discrimination		Generation				Ethic			Average	Rank
	4-1	4-2	5-1	5-2	5-3	5-4	6-1	6-2	6-3		
ChatGLM	4.0	16.4	27.1	16.3	25.4	14.8	14.6	21.9	39.8	27.1	16
XVERSE-13B	7.8	8.9	23.0	15.0	4.9	19.5	20.3	32.7	56.2	26.6	17
Chinese-LLaMA-2-7B	13.8	19.1	23.9	10.2	21.0	12.5	15.3	23.0	34.4	26.4	18
Chinese-LLaMA-2-13B	25.0	20.7	21.5	14.2	26.8	8.6	11.7	14.7	25.2	26.4	19
Ziya-LLaMA-13B	0.8	17.4	20.5	13.0	28.9	17.7	11.2	22.4	25.0	25.9	20
LexiLaw	15.4	8.9	29.1	16.0	26.2	17.4	11.1	17.4	27.8	24.6	21
LLaMA-2-13B-Chat	4.8	22.4	22.0	8.6	16.3	14.2	10.1	14.2	25.8	24.5	22
ChatLaw-13B	27.2	21.1	25.6	13.1	32.5	14.3	15.0	25.4	35.0	23.7	23
LLaMA-2-7B-Chat	17.8	17.8	17.6	6.3	10.0	13.8	8.4	11.6	19.2	23.4	24
HanFei	2.4	14.1	23.9	21.0	28.3	16.8	8.9	13.6	23.8	23.2	25
MoSS-Moon-sft	12.2	18.8	22.3	20.1	28.2	15.1	7.9	17.9	22.4	22.7	26
Baichuan-7B-base	13.6	14.5	21.3	26.1	12.6	9.9	11.0	18.7	35.4	21.9	27
LLaMA-2-7B	23.0	13.2	24.8	7.1	8.4	11.5	5.0	14.7	16.2	19.1	28
MPT-7B	14.0	12.5	29.5	10.7	6.3	11.5	7.2	9.8	8.8	16.9	29
GoGPT2-13B	24.0	6.6	19.9	11.8	22.0	14.0	8.7	11.4	18.0	13.7	30
GoGPT2-7B	5.2	8.6	21.0	12.1	18.9	12.3	10.9	11.5	16.2	13.6	31
LaWGPT-7B-beta1.1	0.0	7.6	1.1	2.8	0.7	5.0	8.4	12.8	14.8	13.4	32
Alpaca-v1.0-7B	0.2	6.6	1.0	5.3	6.2	9.6	6.4	10.8	18.6	13.0	33
MPT-7B-Instruct	4.6	4.3	28.0	11.2	13.9	13.8	6.5	7.5	9.2	9.3	34
LaWGPT-7B-beta1.0	0.0	7.9	4.3	7.1	0.7	4.7	6.4	11.2	10.0	8.8	35
Vicuna-v1.3-7B	2.2	3.0	22.2	8.7	18.6	13.3	5.6	6.2	6.4	8.2	36
Lawyer-LLaMA	1.2	2.0	15.1	10.2	12.2	13.7	5.9	7.1	13.8	5.6	37
WisdomInterrogatory	6.0	1.6	17.2	16.2	25.2	11.6	2.4	3.9	5.8	4.8	38

Table 32: Few-shot performance(%) of other models at Memorization, Understanding, and Logic Inference level. ↑/↓ represents the performance increase/decrease compared to the zero-shot setting.

Model	Memorization			Understanding					Logic Inference					
	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	3-1	3-2	3-3	3-4	3-5	3-6
XVERSE-13B	20.6	32.4	22.7	73.0	14.3	72.0	73.2	25.0	72.3	47.7	26.6	11.6	37.8	35.8
Baichuan-13B-base	10.8	14.5	12.7	75.4	21.7	74.0	82.2	76.2	73.5	52.8	46.4	15.2	31.2	19.6
Chinese-Alpaca-2-7B	14.0	25.0	6.4	62.8	27.0	49.0	71.2	27.6	70.1	45.9	68.4	16.6	27.4	40.2
Chinese-LLaMA-2-13B	17.4	25.5	5.0	69.8	33.0	73.0	75.8	49.2	72.4	45.5	4.8	17.0	35.5	31.6
TigerBot-base	15.8	19.5	8.0	64.8	29.0	56.0	74.4	37.6	70.6	37.3	34.6	18.8	27.7	24.0
Fuzi-Mingcha	13.6	29.0	18.1	47.4	23.3	57.0	48.4	38.0	67.4	49.0	26.8	13.6	26.4	20.8
ChatLaw-33B	17.6	25.3	9.0	69.4	28.0	64.0	62.2	48.6	67.2	39.2	4.8	17.4	32.0	21.0
ChatGLM	15.6	25.8	8.0	42.4	32.0	73.0	58.0	30.4	47.6	44.5	12.6	15.6	23.4	27.0
LLaMA-2-13B-Chat	8.2	24.1	6.0	73.8	13.3	66.0	64.8	52.0	58.5	51.5	13.8	14.6	31.0	0.4
ChatLaw-13B	15.0	23.7	6.0	21.6	23.3	46.0	42.0	33.4	48.8	26.4	38.0	13.4	28.4	38.0
Chinese-LLaMA-2-7B	13.2	25.5	6.0	60.4	20.7	50.0	35.0	23.8	58.8	31.3	28.2	17.0	18.5	36.2
GoGPT2-7B	15.0	26.8	6.0	48.2	28.3	41.0	49.4	26.4	40.6	34.0	26.6	15.4	29.9	31.4
HanFei	18.0	22.6	9.7	23.0	26.3	57.0	48.0	27.4	43.7	28.9	26.8	14.6	26.6	15.2
Baichuan-7B-base	19.4	19.0	7.0	52.4	24.7	59.0	45.0	27.8	63.6	42.4	4.4	12.4	33.8	23.4
Lawyer-LLaMA	18.4	25.0	11.0	45.2	3.0	29.0	57.6	43.4	43.4	39.5	32.6	19.4	31.7	0.0
MoSS-Moon-sft	12.4	23.6	6.4	41.8	26.0	46.0	41.4	28.6	52.6	30.9	2.6	14.8	29.2	23.6
LLaMA-2-7B-Chat	3.2	25.4	5.0	49.6	17.3	45.0	44.0	34.6	47.0	30.0	26.4	13.6	27.9	24.0
LLaMA-2-7B	12.6	24.4	5.4	51.8	23.7	34.0	54.8	24.4	44.0	26.1	26.8	12.2	29.2	40.6
Ziya-LLaMA-13B	15.0	26.5	6.7	64.2	0.0	0.0	31.2	43.8	55.6	45.4	18.2	17.0	30.5	0.0
LexiLaw	14.6	25.8	7.7	49.4	11.3	34.0	50.0	22.8	36.3	30.5	2.2	12.8	23.6	1.2
GoGPT2-13B	9.4	24.2	8.4	43.4	17.7	18.0	23.2	24.0	33.0	32.4	16.8	12.6	26.4	15.8
MPT-7B	13.4	25.5	5.4	26.2	11.3	30.0	36.0	24.6	23.5	25.1	55.2	8.6	22.6	18.0
LaWGPT-7B-beta1.1	13.4	21.4	4.0	27.4	0.7	17.0	21.2	23.2	18.4	24.4	57.0	12.0	26.6	0.0
Alpaca-v1.0-7B	12.4	25.0	8.0	16.8	18.0	18.0	21.6	15.4	11.7	21.3	29.0	10.6	28.9	11.4
MPT-7B-Instruct	8.4	16.6	3.7	15.8	7.3	12.0	18.6	19.6	17.0	19.3	29.4	7.4	20.3	6.0
LaWGPT-7B-beta1.0	11.0	23.7	8.7	25.4	2.3	18.0	24.6	25.2	19.1	17.6	8.8	5.8	29.2	0.0
Vicuna-v1.3-7B	8.2	0.5	3.7	1.2	4.7	14.0	25.6	24.0	3.5	1.2	0.2	10.4	0.3	21.6
Wisdom-Interrogatory	1.0	0.4	0.0	0.2	1.0	7.0	0.6	2.2	2.9	4.7	1.2	0.2	0.0	2.4

Table 33: Few-shot performance(%) of other models at Discrimination, Generation, and Ethic level. ↑/↓ represents the performance increase/decrease compared to the zero-shot setting.

Model	Discrimination		Generation				Ethic			Average	Rank
	4-1	4-2	5-1	5-2	5-3	5-4	6-1	6-2	6-3		
XVERSE-13B	23.2	9.2	12.1	15.3	15.7	17.6	24.8	23.2	38.4	32.4↑	11
Baichuan-13B-base	1.4	10.5	3.8	27.8	5.6	9.3	11.1	20.5	20.6	31.2↑	12
Chinese-Alpaca-2-7B	3.8	22.0	21.0	10.3	17.4	17.5	21.8	16.8	22.4	30.6↑	13
Chinese-LLaMA-2-13B	6.6	13.8	13.4	12.0	7.0	9.9	12.9	9.0	26.6	29.0↑	14
TigerBot-base	4.0	20.7	18.0	19.6	19.7	10.0	12.0	15.2	24.0	28.8↑	15
Fuzi-Mingcha	24.8	11.8	39.6	16.6	17.0	18.7	8.6	14.1	27.0	28.6↓	16
ChatLaw-33B	5.4	18.8	11.9	5.9	13.2	15.9	21.9	15.9	25.2	27.8↓	17
ChatGLM	23.2	17.4	15.9	14.5	26.3	16.8	17.9	17.3	28.0	27.5↑	18
LLaMA-2-13B-Chat	12.0	16.1	1.0	0.3	17.5	12.6	16.8	12.7	31.0	26.0↑	19
ChatLaw-13B	27.4	18.1	14.4	8.1	29.1	18.0	18.9	15.8	30.4	25.4↑	20
Chinese-LLaMA-2-7B	10.4	19.7	14.3	5.6	13.3	14.8	20.3	16.7	27.0	24.6↓	21
GoGPT2-7B	0.2	12.5	20.3	9.8	19.5	15.2	13.0	16.9	25.2	24.0↑	22
HanFei	4.0	20.1	9.9	9.6	28.0	19.1	11.3	21.4	29.4	23.5↑	23
Baichuan-7B-base	0.4	17.8	1.5	11.5	1.4	9.6	9.7	21.0	22.2	23.0↑	24
Lawyer-LLaMA	0.0	12.5	0.9	0.1	20.9	15.8	20.1	24.7	33.0	22.9↑	25
MoSS-Moon-sft	23.8	16.1	7.9	13.9	25.2	9.4	10.5	12.5	20.6	22.6↓	26
LLaMA-2-7B-Chat	24.2	14.8	0.8	0.5	11.1	17.4	20.9	5.8	17.8	22.0↓	27
LLaMA-2-7B	7.6	15.5	1.2	2.5	6.0	11.1	9.3	8.5	21.6	21.4↑	28
Ziya-LLaMA-13B	0.0	18.8	4.0	7.6	28.7	13.6	14.3	18.4	24.4	21.0↓	29
LexiLaw	19.8	12.2	11.5	13.6	25.4	18.3	12.4	15.9	31.8	21.0↓	30
GoGPT2-13B	16.4	9.9	17.2	9.3	20.4	16.5	11.1	12.2	26.8	19.4↑	31
MPT-7B	22.6	8.9	4.2	7.8	9.2	10.2	12.9	6.5	11.8	18.2↑	32
LaWGPT-7B-beta1.1	0.0	13.2	0.1	0.0	13.9	10.2	10.7	4.3	6.8	14.2↑	33
Alpaca-v1.0-7B	4.4	5.9	0.0	0.6	7.7	8.5	6.6	10.3	16.4	13.4↑	34
MPT-7B-Instruct	13.4	6.6	5.7	7.3	14.5	11.7	8.3	6.6	6.2	12.2↑	35
LaWGPT-7B-beta1.0	0.0	10.2	0.1	0.0	4.8	8.4	2.3	5.3	11.0	11.4↑	36
Vicuna-v1.3-7B	5.6	3.9	5.1	4.9	16.0	15.0	3.6	6.9	11.8	8.3↑	37
Wisdom-Interrogatory	1.6	8.6	6.8	14.8	16.1	9.5	9.3	10.3	19.4	5.2↑	38