# CitaLaw: Enhancing LLM with Citations in Legal Domain

**Kepu Zhang[1], Weijie Yu[2]*, Sunhao Dai[1], Jun Xu[1]**
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2] University of International Business and Economics
kepuzhang@ruc.edu.cn, yu@uibe.edu.cn

## Abstract

In this paper, we propose CitaLaw, the first benchmark designed to evaluate LLMs' ability to produce legally sound responses with appropriate citations. CitaLaw features a diverse set of legal questions for both laypersons and practitioners, paired with a comprehensive corpus of law articles and precedent cases as a reference pool. This framework enables LLM-based systems to retrieve supporting citations from the reference corpus and align these citations with the corresponding sentences in their responses. Moreover, we introduce syllogism-inspired evaluation methods to assess the legal alignment between retrieved references and LLM-generated responses, as well as their consistency with user questions. Extensive experiments on 2 open-domain and 7 legal-specific LLMs demonstrate that integrating legal references substantially enhances response quality. Furthermore, our proposed syllogism-based evaluation method exhibits strong agreement with human judgments.

## 1 Introduction

Generating responses supported by citations, such as relevant law articles and precedent cases, is essential for ensuring the trustworthiness of large language models (LLMs) in legal tasks. For laypersons seeking legal advice (Fei et al., 2023), LLM-generated responses grounded in citations provide verifiable information, fostering trust in the system. Conversely, for legal practitioners such as lawyers and judges, citations serve as supportive evidence that aids in analyzing complex cases, validating legal arguments, and ensuring decisions align with established legal principles (Li et al., 2024; Zhong et al., 2020; Abdallah et al., 2023).

Recently, a growing body of benchmark research (Gao et al., 2023a; Li et al., 2023) has focused on enabling LLMs to provide citations for the statements they generate. For instance, ALCE (Gao et al., 2023b) introduces a benchmark designed to evaluate the ability of LLMs to generate citation-supported outputs, aiming to improve factual accuracy. WebCiteS (Deng et al., 2024) provides a curated database of manually annotated summaries and citations to enhance performance in text summarization and citation generation.

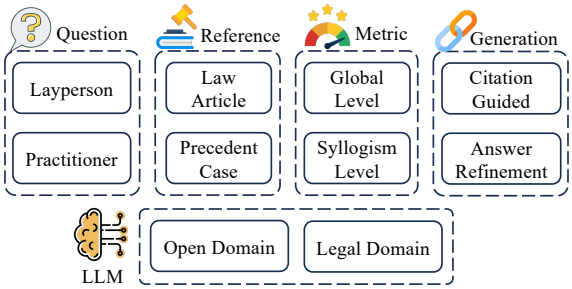



Figure 1: The framework of our CitaLaw.

While these studies have made notable progress in general domains, they face significant challenges when applied to the legal domain. **First**, laypersons and legal practitioners interact with LLMs differently and have distinct expectations for citations. Laypersons typically seek legal advice and rely on citations to verify the accuracy of LLM responses, whereas legal practitioners pose more complex queries, using LLMs for legal reasoning, with citations serving as supportive evidence. Existing studies fail to address these differences, leading to unsatisfactory performance in real-world applications. **Second**, existing methods often fall short in providing the diverse references required in legal contexts, such as law articles and precedent cases. Law articles establish the foundational legal framework, while precedent cases offer concrete examples and interpretive guidance. These two types of references inherently align with the distinct characteristics of civil and common law systems. **Third**, traditional citation evaluation measures, such as ROUGE (Lin, 2004), rely on surface-level similar-

* Corresponding author



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 11183–11196

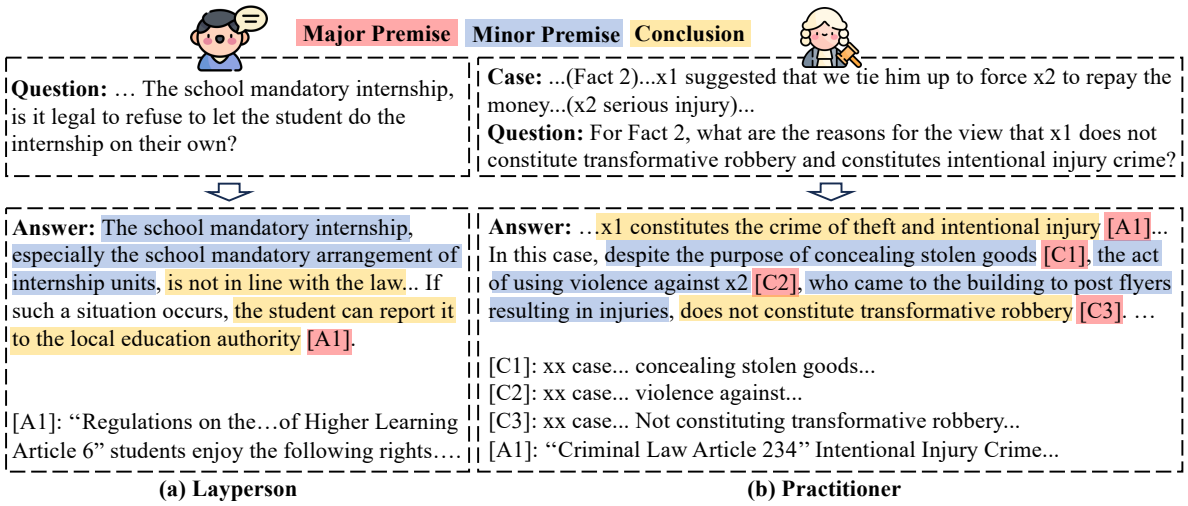



Figure 2: Examples from the two subsets of CitaLaw, with text in red, blue, and yellow representing the three dimensions of the syllogism: major premise, minor premise (circumstances, illegal acts), and conclusion (legal decisions), respectively. [A] and [C] denote citations to relevant law articles and precedent cases, respectively.

ities and are often insufficient to assess the alignment between references and LLM-generated responses. In the legal domain, effective evaluation requires a deeper understanding of logical and semantic relationships.

To overcome the above challenges, we propose CitaLaw, the **first** benchmark tailored to evaluate LLMs' capabilities in generating legally grounded responses supported by accurate and context-aware citations. As shown in Figure 1, CitaLaw incorporates four distinct legal-specific features:

(1) CitaLaw has two subsets tailored for laypersons and practitioners, with examples in Figure 2. Laypersons typically ask shorter, conversational questions, while practitioners often pose specialized, detailed questions.

(2) CitaLaw includes a retrieval corpus comprising two commonly used references: law articles, which provide clear and concise guidelines for addressing user questions, and precedent cases, which offer legal reasoning and support for judicial decisions. Recognizing the distinct needs of laypersons and practitioners, we provide only law articles for laypersons to ensure clarity, while practitioners have access to both law articles and precedent cases to support more complex legal reasoning.

(3) In addition to traditional global-level metrics such as MAUVE (Pillutla et al., 2021), we propose a syllogism-based evaluation method to assess both the response correctness and the citation quality. This method provides a more granular evaluation by focusing on three key dimensions: circumstances, illegal acts, and legal decisions.

(4) We consider two types of response generation methods. The first type, Citation-Guided Generation (CGG), involves generating responses by incorporating retrieved references during generation. The second type, Answer Refinement Generation (ARG), refines the LLMs' initial response (CloseBook) by retrieving and incorporating reference information. This category includes ARG-Q, which retrieves citations using only the user query, and ARG-QA, which retrieves citations using both the user query and the LLM's initial response.

Extensive experiments on two open-domain and seven legal-specific LLMs reveal the following key insights: 1) Incorporating legal references into the LLM significantly improves the quality of responses; 2) Including references as part of the LLM's input consistently outperforms answer-refinement methods; 3) Leveraging references to refine the LLM's responses yields better alignment of responses and references. 4) For fine-tuning LLMs in legal scenarios, incorporating law articles, syllogistic reasoning, and full-scale fine-tuning achieves promising performance. 5) Open-domain LLMs surprisingly outperform legal-specific LLMs in certain scenarios; 6) Human evaluations show a strong correlation with our syllogism-based methods.

In summary, our contributions are as follows:

- To the best of our knowledge, CitaLaw is the **first** benchmark designed to evaluate the capability of LLMs to generate legally grounded responses with accurate and context-aware citations. CitaLaw includes questions tailored to both laypersons and practitioners, paired

with a citation corpus comprising law articles and precedent cases.

- We propose a two-level evaluation framework that combines global-level metrics with a syllogism-based reasoning approach. Additionally, we explore two mainstream methods for legal response generation: citation-guided and answer refinement.

- Through extensive experiments on two open-domain and seven legal-specific LLMs, we demonstrate the effectiveness of integrating legal references into response generation and validate our syllogism-based evaluation method. Additionally, we provide actionable insights for the practical deployment of LLMs in legal scenarios.

## 2 Related Work

**LLM for Legal Task.** A amount of work has explored applying LLMs to legal tasks (Savelka et al., 2023; Wu et al., 2023b; Yu et al., 2022a; Blair-Stanek et al., 2023). Building LLMs tailored for legal scenarios is a popular direction (Yue et al., 2023; Wu et al., 2023a; He et al., 2023). There are also some benchmarks that explore the capabilities of LLMs in legal tasks. LawBench (Fei et al., 2023) evaluates LLMs' legal knowledge across three cognitive aspects. LAiW (Dai et al., 2023) assesses LLMs' legal reasoning abilities based on legal practice logic. LexEval (Li et al., 2024) evaluates LLMs' legal capabilities based on a new legal cognitive ability classification system. However, none of them have considered enhancing the trustworthiness of LLMs in legal scenarios by generating outputs with citations.

**Citation in LLM.** Attribution (Li et al., 2023) in LLMs refers to providing supporting evidence for the answers generated by the model, presented in the form of citations. ALCE (Gao et al., 2023b) is an automated benchmark for evaluating LLMs' ability to generate outputs with citations, aimed at improving the factual accuracy of the generated responses. WebCiteS (Deng et al., 2024) provides a database containing 7,000 manually annotated summaries and citations to enhance LLMs' capabilities in summarization and citation. RARR (Gao et al., 2023a) enhances LLM outputs by automatically adding citations, and modifying the responses. ExpertQA (Malaviya et al., 2024) verifies and modifies citations through expert review to ensure reliability. In contrast to the above works, `CitaLaw` focuses specifically on citation in legal scenarios.

## 3 Task Setup and Dataset Construction

Suppose we have a legal corpus $D$, which consists of either a collection of precedent cases ($D_l$) or law articles ($D_c$). Given a user question $x$ posed by either a layperson or a practitioner, the LLM-based system is tasked with retrieving supportive citations from $D$ and generating a legally grounded response $y$. The response $y$ comprises a list of $n$ sentences, i.e., $y = [s_1, \cdots, s_n]$, where each sentence $s_i$ refers to at most one corresponding citation. As illustrated in Figure 2, the system is further required to attach each citation to its relevant sentence, with "[A]" and "[C]" denoting references to law articles and precedent cases, respectively.

To enable the evaluation of this task, we construct the specialized dataset (Table 1 shows the statistics) as follows:

To simulate the behavior of **laypersons**, we include questions that are more conversational, lack detailed case descriptions, and are relatively short in length. We use the consultation section from LawBench (Fei et al., 2023), which collects user queries from the Hualv website[1] and answers provided by lawyers or legal consulting firms.

To simulate the behavior of **legal practitioners**, we include questions that are more professional, often accompanied by detailed case descriptions, and generally longer. For this purpose, we use the open-ended question section from LexEval (Li et al., 2024), which consists of subjective questions from the National Uniform Legal Profession Qualification Examination. These questions are particularly challenging for LLMs, requiring them to understand the case fully and apply legal knowledge accurately to generate answers.

In terms of the **corpus**, we construct a comprehensive corpus from multiple sources, including law articles and precedent cases. Specifically, for law articles, we collect approximately 50,000 documents from LexiLaw[2], covering areas such as Civil Law, Criminal Law, and judicial interpretations. For precedent cases, we include both criminal and civil cases. Criminal cases are sourced from the LeCaRD legal retrieval dataset (Ma et al., 2021b), ELAM (Yu et al., 2022b), and civil cases from the CAIL legal summary

[1] www.66law.com

[2] https://github.com/CSHaitao/LexiLaw

| Dataset | #Q | $\text{Len}_Q$ | $\text{Len}_A$ | Q Type |
|---|---|---|---|---|
| Layperson | 500 | 57.62 | 107.40 | Question |
| Practitioner | 500 | 618.96 | 193.46 | Case + Question |

Table 1: Dataset statistics. #Q indicates the number of questions, $\text{Len}_Q$ and $\text{Len}_A$ denote the average lengths of questions and gold answers, and Q Type refers to the question type.

dataset, LJP-MSJudge (Ma et al., 2021a), and the pre-training data of fuzi.mingcha (Wu et al., 2023a). As a supplement to precedent cases, we also incorporate question-and-answer pairs from fine-tuning datasets of legal LLMs as part of the precedent cases. These QA pairs are collected from DISC-LawLLM (Yue et al., 2023), LawGPT_zh (Liu et al., 2023), and HanFei (He et al., 2023). In total, the constructed corpus contains approximately 500,000 documents, ensuring sufficient coverage of both law articles and precedent cases to support diverse legal tasks.

# 4 Method

## 4.1 Response Generation

We consider two types of methods in this study.

**Citation-Guided Generation (CGG)** produces response $y_{cgg}$ given a user question $x$ by referring retrieved relevant document(s) $D_R$:

$$y_{\text{cgg}} = f_{\text{LLM}}(x, D_R, p_1), \tag{1}$$

where $f_{\text{LLM}}$ denotes a open-domain or a legal specific LLM; $p_1$ is the direct generation prompt. All prompt settings are detailed in Appendix A.

**Answer Refinement Generation (ARG)** is a two-stage method that generates the final response $y_{\text{arg}}$ by refining the LLM's initial response $y_{\text{init}}$ through the retrieval and incorporation of reference information. This process can be formulated as:

$$y_{\text{init}} = f_{\text{LLM}}(x, p_2), \tag{2}$$

where $p_2$ is the prompt instructing the LLM to directly generate an initial response without reference information. We refer to this step as **CloseBook**. The initial response $y_{\text{init}}$ is then refined as:

$$y_{\text{arg}} = f_{\text{LLM}}(y_{\text{init}}, D_R, p_3), \tag{3}$$

where $p_3$ is the prompt guiding the LLM to refine the $y_{\text{init}}$ using the retrieved documents $D_R$.

Laypersons and practitioners interact with LLMs differently and have distinct expectations for citations. When $x$ is submitted by a layperson, the corresponding $D_R$ consists of relevant law articles. In contrast, when $x$ is submitted by a practitioner, the corresponding $D_R$ includes both relevant law articles and precedent cases. The process for retrieving $D_R$ from $D$ is detailed in the next subsection.

## 4.2 Citation Retrieval

We explore state-of-the-art open-domain dense retriever BGE (Xiao et al., 2023), along with two legal-specific dense retrievers, Criminal-BERT (Zhong et al., 2019) and Civil-BERT (Zhong et al., 2019). We also investigate two types of retrieval queries: $x$ (the user question alone, **ARG-Q**) and $[x; y_{\text{init}}]$ (the concatenation of the user query $x$ and the initial response $y_{\text{init}}$, where [;] denotes the concatenation operation, **ARG-QA**). The impact of different retrieval models on performance will be analyzed in the experiments.

## 4.3 Citation Attachment

Building on the retrieved citations, this subsection outlines the process of attaching these law articles or precedents to specific sentences in the LLM-generated responses. This process involves answering two key questions:

**What kind of sentences can be associated with citations?** We utilize co-occurring words and legal entity extraction to identify sentences that explicitly reference legal concepts, actions, or terms relevant to the retrieved citations. Specifically, we construct a pool of legal terminologies using THUOCL[3] and LaWGPT (Zhou et al., 2024). A sentence is considered eligible if it contains any of the terminologies from this pool. Additionally, we use SpaCy (Honnibal et al., 2020) to extract legal entities from each sentence. If a sentence includes legal entities, it is also deemed eligible for citation attachment.

**How are citations attached to the identified sentences?** If a sentence is deemed eligible for citation attachment, we associate it with retrieved citations as follows. For the laypersons, the retrieved law article $c_l \in D_l$ is attached to the most relevant sentence $s_k \in y$:

$$C_{\text{Lay}} = \{(s_k, c_l) \mid s_k = \arg\max_{s_i \in y} \text{sim}(s_i, c_l)\}, \tag{4}$$

where $(s_k, c_l)$ represents attaching the reference $c_l$ to the sentence $s_i$, and $\text{sim}(\cdot)$ is computed using sentence-BERT (Reimers, 2019). We set $|C_{\text{Lay}}| = 1$ because, typically, a layperson's query pertains to only one specific legal article. For practitioners,

[3]https://github.com/thunlp/THUOCL

we attach the retrieved law article in the same way as for laypersons. Additionally, we associate the retrieved precedent cases $c_c \in D_c$ with each $s_i \in y$, which is formulated as:

$$\begin{aligned} C_{\text{Pra}} =& \{(s_k, c_l) \mid s_k = \arg\max_{s_i \in y} \text{sim}(s_i, c_l)\} \\ & \cup \{(s_i, c_c)|, c_c = \arg\max_{c_j \in D_c} \text{sim}(s_i, c_j)\}, \end{aligned} \tag{5}$$

where $|D_c| = 3$, meaning each response $y$ can be associated with up to three precedents[4].

## 5 Evaluation

CitaLaw provides a comprehensive evaluation framework incorporating metrics for fluency, correctness, and citation quality. This framework is divided into two levels of analysis: global level and the proposed syllogism level.

Syllogism, a foundational framework in legal reasoning, comprises three key components: the major premise, the minor premise, and the conclusion. In our legal context, these correspond to the relevant law article or precedent case (major premise), the factual circumstances and actions of a specific case (minor premise), and the resulting legal decision (conclusion). By integrating this syllogistic framework, CitaLaw goes beyond surface-level correctness to evaluate the logical coherence and alignment of LLM-generated responses with established legal principles.

### 5.1 Fluency (Style Consistency)

To ensure the LLM-generated responses align with the user's requirements, the system must adapt its style based on the user's background. For laypersons, responses should avoid excessive technical jargon to ensure accessibility and comprehension. Conversely, responses for legal practitioners should adopt a formal and professional tone to maintain credibility and utility. To achieve this aim, we concatenate the user query and the LLM-generated response and apply MAUVE (Pillutla et al., 2021) to assess their style consistency.

### 5.2 Correctness

At the **global level**, we use established metrics ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019). ROUGE measures word-level overlap between the generated and labeled responses, with scores reported for ROUGE-1, ROUGE-2, and ROUGE-L. BERTScore captures semantic similarity between the generated and labeled responses, and we report the F-score (BERT-F) for evaluation. These metrics assess the overall correctness of LLM-generated responses.

At the **syllogism level**, we leverage the Qwen2 (Yang et al., 2024) to extract key components, including the circumstances, illegal acts, and legal decisions. We use sentence-BERT (Reimers, 2019) to measure the alignment between the labeled responses and the generated outputs across these dimensions, resulting in $\text{Correct}_\text{c}$, $\text{Correct}_\text{a}$, and $\text{Correct}_\text{d}$. This syllogism-level evaluation allows us to assess the logical coherence of the responses, ensuring that they align with the underlying legal reasoning principles.

### 5.3 Citation Quality

As previously discussed, we assume that a question submitted by laypersons typically corresponds to a specific law article. Therefore, at the **global level**, we evaluate the citation quality of the retrieved law article (premise) by measuring its entailment with the associated sentence in the LLM's response (hypothesis). Specifically, we use an NLI model to compute $\text{Cita}_\text{Law}$, which quantifies the degree to which the law article entails the attached sentence. This metric reflects how effectively the response aligns with the cited law article. We employ DISC-LawLLM (Yue et al., 2023) as the NLI model due to its strong agreement with human evaluations (as discussed in Sec. 6.3) and its superior performance compared to other NLI models (as detailed in Sec. 6.5).

At the **syllogism level**, we evaluate the quality of precedent case citations by examining three key components: circumstances, illegal acts, and legal decisions. After extracting these elements from both the retrieved cases and the associated sentence in the LLM's response, we utilize DISC-LawLLM to assess the entailment for each component. This evaluation yields three distinct scores: $\text{Cita}_\text{c}$ for circumstances, $\text{Cita}_\text{a}$ for illegal acts, and $\text{Cita}_\text{d}$ for legal decisions, providing a more detailed and nuanced assessment of citation quality within the syllogism framework.

## 6 Experiments

We conduct extensive experiments on our CitaLaw using the proposed two-level evaluation methods.

[4]Considering the input window size of LLMs, we set up to retrieve 3 precedent cases.

| Metric | | Fluency | Correctness | | | | | | | Citation | All |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Model | Mauve | Rouge-1 | Rouge-2 | Rouge-L | BERT-F | $\text{Correct}_\text{c}$ | $\text{Correct}_\text{a}$ | $\text{Correct}_\text{d}$ | $\text{Cita}_\text{Law}$ | **Avg** |
| Llama3 (Llam3-8B-Instruct) | CloseBook | 22.63 | 16.47 | 1.95 | 13.34 | 58.46 | **73.05** | 68.24 | 66.87 | 67.38 | 43.15 |
| | CGG | 61.01 | **23.97** | 6.05 | 17.91 | **65.94** | 67.29 | **77.31** | **74.95** | **86.70** | **53.46** |
| | ARG-Q | **61.27** | 23.17 | 5.65 | 17.83 | 64.23 | 69.04 | 75.45 | 74.47 | 79.10 | 52.24 |
| | ARG-QA | 51.83 | 23.73 | **6.96** | **18.53** | 64.84 | 71.37 | 74.81 | 74.66 | 80.80 | 51.95 |
| Qwen2 (Qwen2-7B-Instruct) | CloseBook | 21.04 | 15.29 | 2.27 | 11.31 | 58.39 | **70.89** | 71.71 | 69.93 | 72.35 | 43.69 |
| | CGG | **75.10** | **22.26** | 4.77 | 15.41 | **65.28** | 67.50 | **78.62** | **77.82** | 77.59 | **53.82** |
| | ARG-Q | 66.55 | 20.86 | 4.50 | 15.42 | 64.59 | 66.96 | 77.82 | 75.66 | 81.48 | 52.65 |
| | ARG-QA | 66.80 | 21.73 | **4.78** | **16.34** | 64.85 | 69.31 | 76.35 | 75.05 | **82.83** | 53.11 |
| Legal LLM (CGG) | DISC-LawLLM | **72.70** | 22.46 | 4.14 | 15.48 | 65.06 | 65.21 | 78.55 | 76.17 | 83.46 | 53.69 |
| | fuzi.mingcha | 56.58 | 24.54 | 5.70 | 17.48 | **65.86** | 63.28 | **79.56** | **77.94** | 81.64 | 52.51 |
| | LexiLaw | 71.89 | **24.96** | **6.25** | **18.91** | 65.68 | 68.89 | 78.12 | 76.72 | 82.42 | **54.87** |
| | Tailing | 13.95 | 15.93 | 4.13 | 12.89 | 59.47 | **72.00** | 69.11 | 68.38 | 82.67 | 44.28 |
| | zhihai | 37.50 | 20.98 | 4.59 | 13.69 | 64.54 | 67.75 | 77.68 | 76.99 | 77.16 | 48.99 |
| | LawGPT_zh | 51.60 | 23.33 | 5.28 | 16.17 | 65.14 | 63.72 | 79.43 | 77.52 | **86.18** | 52.04 |
| | Hanfei | 51.12 | 23.95 | 5.19 | 18.76 | 65.12 | 70.83 | 75.01 | 74.21 | 76.97 | 51.24 |

Table 2: Performance comparisons on the Layperson dataset. The best performance is indicated in bold.

## 6.1 Experimental Settings

### 6.1.1 Evaluated Models

We selected two categories of LLMs for testing: The legal LLMs include (1) **fuzi.mingcha** (6B) (Wu et al., 2023a), (2) **LexiLaw**[5] (6B), (3) **Tailing**[6] (7B), (4) **DISC-LawLLM** (13B) (Yue et al., 2023), (5) **zhihai** (7B) (Wu et al.), (6) **LawGPT_zh** (6B) (Liu et al., 2023), (7) **HanFei** (7B) (He et al., 2023). The open-domain LLMs include Qwen2 (7B) (Yang et al., 2024) and Llama3 (8B) (AI@Meta, 2024). For these models, we tested all methods mentioned in Sec. 4, including: (1) **CloseBook**, (2) **CGG**, (3) **ARG-Q** and (4) **ARG-QA**. For the legal LLMs, we generate responses using CGG. Appendix B has the details.

### 6.1.2 Implementation Details

Our implementation is based on the Huggingface Transformers library (Wolf et al., 2020) with PyTorch. We use bge-base-zh-v1.5 (Xiao et al., 2023) as the retrieval model and conduct all experiments on Nvidia A6000 GPUs. Additional details are provided in Appendix C and `https://github.com/ke-01/CitaLaw`.

## 6.2 Main Results

The results on the Layperson and Practitioner datasets are presented in Table 2 and Table 3. We analyze the results from three perspectives:

### 6.2.1 Peformance of Open-Domain LLM

**Legal references improve the response quality.** Compared to CloseBook, the overall performance in CGG, ARG-Q, and ARG-QA has improved. This indicates that incorporating references into the LLM helps it better understand both the question and the required direction for the answer, thereby enhancing performance in terms of style consistency, correctness, and citation quality.

**CGG achieves better response quality.** We observe that CGG achieves optimal performance, especially response correctness, suggesting that incorporating legal references into the LLM input is more effective than refining the LLM's response. This is because including legal knowledge as input allows the LLM to consider relevant context when generating replies, whereas refining the response might lead to excessive alterations.

**ARG improves the alignment of responses and references.** We can observe that ARG outperforms CGG in citation-related metrics overall. This is because CGG merely incorporates reference information as input, which may lead the model to overlook some reference details during the generation process. In contrast, ARG modifies the answer based on the references after generation, making it easier to ensure the completeness of citations.

**Chinese data fine-tuning can bring benefits.** Both the Layperson and Practitioner datasets are Chinese datasets. Qwen2 (Fine-tuning on more Chinese data) achieved better performance than Llama3, demonstrating the benefits of using Chinese data for fine-tuning.

**CloseBook tends to state circumstances.** CloseBook performs better in terms of correctness regarding circumstances compared to the other dimensions. This suggests that when judicial knowledge references are not used, the LLM is more likely to repeat the circumstances itself, rather than providing an appropriate response to the illegal acts and the legal decision.

[5]https://github.com/CSHaitao/LexiLaw
[6]https://github.com/DUTIR-LegalIntelligence/Tailing

| Metric | | Fluency | Correctness | | | | | | | Citation | | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Model | Mauve | Rouge-1 | Rouge-2 | Rouge-L | BERT-F | $\text{Correct}_c$ | $\text{Correct}_a$ | $\text{Correct}_d$ | $\text{Cita}_{\text{Law}}$ | $\text{Cita}_c$ | $\text{Cita}_a$ | $\text{Cita}_d$ | **Avg** |
| Llama3 (Llam3-8B-Instruct) | CloseBook | 23.81 | 23.05 | 7.29 | 19.23 | 62.83 | **76.30** | 71.05 | 70.32 | 63.49 | 66.95 | 68.83 | 65.46 | 51.55 |
| | CGG | 36.37 | **26.15** | **7.84** | **19.55** | **65.60** | 67.19 | **76.36** | **77.73** | **73.58** | 68.23 | 67.87 | 67.65 | **54.51** |
| | ARG-Q | **42.65** | 20.39 | 5.07 | 15.75 | 62.82 | 70.49 | 73.67 | 72.00 | 68.61 | **69.48** | **70.51** | 68.34 | 53.31 |
| | ARG-QA | 36.94 | 18.64 | 4.56 | 14.63 | 61.50 | 71.07 | 72.38 | 70.32 | 69.40 | 68.95 | 70.42 | **69.51** | 52.36 |
| Qwen2 (Qwen2-7B-Instruct) | CloseBook | **61.91** | 30.44 | 10.54 | **23.53** | 67.55 | **74.35** | 79.84 | 78.52 | 68.55 | 68.03 | 70.30 | 69.71 | **58.61** |
| | CGG | 39.66 | **31.01** | **10.75** | 23.43 | **69.06** | 73.49 | **80.11** | **81.11** | 70.37 | 67.82 | 69.53 | 70.01 | 57.20 |
| | ARG-Q | 41.02 | 20.57 | 5.14 | 15.62 | 63.31 | 67.84 | 74.71 | 73.94 | **73.01** | 68.96 | **73.20** | **73.64** | 54.25 |
| | ARG-QA | 21.97 | 16.67 | 3.06 | 12.47 | 60.70 | 67.49 | 71.16 | 70.88 | 71.76 | **69.01** | 71.04 | 71.33 | 50.63 |
| Legal LLM (CGG ) | DISC-LawLLM | 38.11 | 21.37 | 6.75 | 16.96 | 60.84 | **73.42** | 72.14 | 71.79 | 63.92 | 67.42 | 68.22 | 65.45 | 52.20 |
| | fuzi.mingcha | 66.55 | 28.95 | 9.51 | 22.69 | 67.06 | 70.73 | 76.66 | 77.47 | 65.92 | 66.94 | **69.28** | 68.69 | 57.54 |
| | LexiLaw | 57.74 | 29.01 | 8.93 | 23.83 | 65.63 | 70.36 | 76.67 | 75.97 | 65.28 | 66.93 | 68.89 | 68.03 | 56.44 |
| | Tailing | 50.16 | 26.52 | 9.16 | 22.44 | 65.35 | 75.96 | 73.83 | 70.30 | 64.65 | 66.94 | 67.56 | 66.09 | 54.91 |
| | zhihai | 26.29 | 21.38 | 6.00 | 15.53 | 64.47 | 65.59 | 76.38 | 77.37 | 67.93 | 66.30 | 63.17 | 59.82 | 50.85 |
| | LawGPT_zh | 47.10 | 29.16 | 8.92 | 22.55 | 67.64 | 69.48 | **79.37** | **80.23** | 66.90 | **68.38** | 67.55 | **68.94** | 56.35 |
| | HanFei | **75.72** | **32.98** | **12.46** | **26.91** | **68.72** | 73.25 | 78.63 | 78.11 | **67.03** | 67.45 | 68.63 | 67.73 | **59.80** |

Table 3: Performance comparisons on the Practitioner dataset. The best performance is indicated in bold.

#### 6.2.2 Performance of Legal LLM

**Law article training achieves gains.** In the Layperson dataset, LexiLaw achieves optimal performance overall. This is because the questions in the Layperson dataset often require only law articles to provide answers clearly, and LexiLaw's training explicitly used law articles, allowing it to effectively handle such questions.

**Full-parameter training offers advantages.** Hanfei achieves the best results in the Practitioner dataset, as it is a fully parameter-trained legal LLM. Full-parameter fine-tuning allows it to effectively simulate a legal expert, thus performing well.

**Syllogistic reasoning is useful.** fuzi.mingcha performs well on syllogism evaluation metrics, particularly on the Layperson dataset. This is due to its fine-tuning of syllogism judgment data.

#### 6.2.3 Open Domain LLM vs. Legal LLM

**Impact of LLM Backbone.** We can observe that some legal LLMs perform worse than open-domain LLMs. This is because Qwen2 and Llama3 are the latest open-domain LLMs, and their overall capabilities have significantly improved. In contrast, most legal LLMs are built on earlier generations of LLMs, which have weaker base models, leading to poorer overall performance.

**Effectiveness of legal knowledge.** Overall, the upper limit of legal LLMs is higher than that of open-domain LLMs. This is because legal LLMs, after extensive training on legal knowledge, have developed strong capabilities in solving legal issues. As a result, even though their base models are outdated, they can still perform effectively.

### 6.3 Human Evaluation

In this section, we compared the syllogism-level metric with human evaluation. Details of legal human annotators can be found in Appendix D. The syllogism-level evaluation of citation quality is divided into two stages: Stage 1: Extracting key components. Stage 2: Assessing the entailment using an NLI model.

**Stage 1:** We randomly selected 50 questions each from the Layperson and Practitioner datasets. After splitting the cases into individual clauses, annotators were provided with the full case and its clauses. They do a three-class classification of each clause. The Qwen2's annotations were then compared with human annotations. The Cohen's kappa coefficient (Cohen, 1960) of 0.7876 indicates substantial agreement (0.61–0.80) between the model's and human annotators' labels.

**Stage 2:** We randomly selected 50 questions from the Practitioner dataset and used Qwen2 to extract key components of pairs of responses and citations. Annotators assessed the degree to which the citations entailed the corresponding response components using a 5-point scale (1: low, 5: high), with descriptions provided in Appendix D. The entailment probabilities given by DISC-LawLLM, which range from 0 to 1, were scaled to the same 1–5 range by multiplying by 5 and rounding. We then compared the scaled model outputs with the human evaluations and calculated Cohen's kappa coefficient. The kappa score of 0.6923 again indicates substantial agreement (0.61–0.80) between the model and human judgments.

### 6.4 Effects on Different Retrieval Models

We selected BGE as the retrieval model in the main experiment. In this section, we explore the impact of using different retrieval models. Specifically, we evaluate Criminal-BERT (Zhong et al., 2019) and Civil-BERT (Zhong et al., 2019), two legal domain models based on BERT, fine-tuned on large-scale criminal and civil law documents, respectively. We replaced the retrieval model and tested the CGG

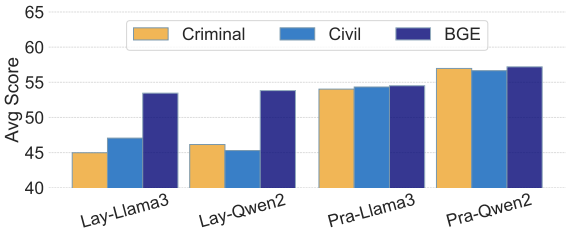
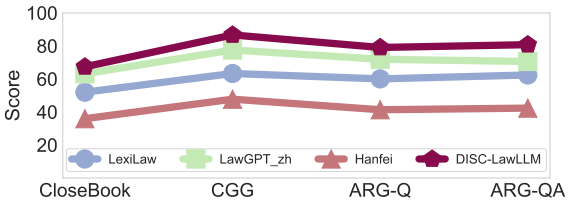
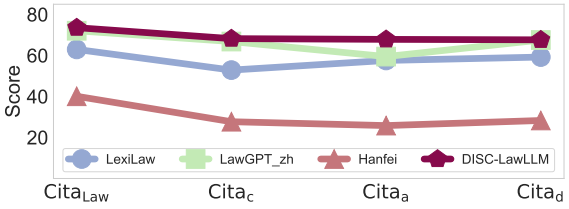



Figure 3: Performance of different retrieval models. Lay is short for Layperson dataset and Pra is short for Practitioner dataset.




(a) Methods for $\text{Cita}_{\text{Law}}$ metric with Layperson dataset.




(b) Metrics for CGG method with Practitioner dataset.

Figure 4: The performance of different NLI models when the LLM is Llama.

method on the Layperson dataset. The average results across all metrics are shown in Figure 3, with detailed metric results provided in Appendix E.

As shown, on the Layperson dataset, BGE significantly outperforms the other two models. This is because the dataset consists of questions from laypersons, which are more everyday in nature. In contrast, the two legal BERT models, having been trained extensively on legal cases, show a distributional mismatch with open-domain data, leading to poorer performance. On the Practitioner dataset, which features professional legal questions, BGE still achieves the best performance. This can be attributed to its extensive training on diverse data, likely including some legal data, and its use of more advanced model architectures and techniques. However, the two legal BERT models perform comparably to BGE, showcasing the benefits of their specialized training on legal data.

### 6.5 Effects on Different NLI Models

We opted to use legal LLMs as the NLI model in our experiments, as they support longer input lengths and incorporate substantial legal knowledge. In Section 6.3, we verified that DISC-LawLLM and human achieved good consistency. In this section, we explore the performance of several legal LLMs in the NLI task. Besides DISC-LawLLM, we evaluated LexiLaw, LawGPT_zh, and Hanfei, which demonstrated strong performance in the main experiments.

In Figures 4 (a), we examined the ability of four legal LLMs to evaluate Llama across the CloseBook, CGG, ARG-Q, and ARG-QA methods using the $\text{Cita}_{\text{Law}}$ metric on the Layperson dataset. In Figures 4 (b), we investigated the performance of four legal LLMs in evaluating the CGG method applied to Llama across the metrics $\text{Cita}_{\text{Law}}$, $\text{Cita}_{\text{c}}$, $\text{Cita}_{\text{a}}$, and $\text{Cita}_{\text{d}}$ on the Practitioner dataset.

We can observe that Hanfei provides lower entailment scores across both datasets. This is because it is a fully parameter-tuned legal LLM, which results in a diminished capability to handle the general task of entailment reasoning. Additionally, we found that on the Practitioner dataset, other legal LLMs achieved results closer to those of DISC-LawLLM, while on the Layperson dataset, the performance gap was significantly larger. This is because the Practitioner dataset is more judicially oriented, aligning with the knowledge seen during the fine-tuning of legal LLMs. In contrast, due to limited training on general-purpose data, other legal LLMs struggle to accurately determine entailment relationships in the Layperson dataset. Similar conclusions can be drawn when the LLM is Qwen in Appendix F.

## 7 Conclusion

We introduce `CitaLaw`, a benchmark designed to explore LLMs to generate responses with citations in legal scenarios, thus improving the trustworthiness of LLMs. `CitaLaw` includes two categories of questions: laypersons and practitioners. For laypersons, `CitaLaw` provides law articles as citations to help them understand the LLM's response clearly. For practitioners, both law articles and precedent cases are provided as citations, better supporting their needs for complex reasoning. `CitaLaw` offers global-level and syllogism-level metrics and supports the integration of citations into LLM inputs to guide generation or using citations to refine LLM's response. We conducted extensive experiments on 7 legal-domain LLMs and 2 popular open-domain LLMs, providing valuable insights for the deployment of LLMs in legal scenarios.

## 8 Limitations

While CitaLaw provides a robust framework for evaluating LLMs in legal scenarios, several limitations should be acknowledged to guide future extensions of this work.

First, the datasets used in CitaLaw are primarily sourced from the Chinese legal system, which may limit the benchmark's applicability to other jurisdictions. However, by incorporating both law articles and precedent cases to align with the principles of civil and common law systems, CitaLaw demonstrates strong potential for adaptation to diverse legal contexts.

Second, the syllogism-based evaluation framework simplifies legal reasoning into three key components: the major premise (law articles or precedent cases), the minor premise (case circumstances and actions), and the conclusion (legal decision). While this structured approach is effective for systematic evaluation, real-world legal reasoning may encompass additional complexities.

## 9 Ethical Considerations

**Data Privacy and Confidentiality.** The legal datasets used in CitaLaw include law articles, precedent cases, user questions, and golden responses. These documents were sourced from publicly available databases, ensuring compliance with data privacy and confidentiality standards. We carefully reviewed the datasets to ensure that no personally identifiable information (PII) or sensitive details about individuals were inadvertently included.

**Alignment with Legal Standards.** Legal AI systems must align with the ethical and professional standards of the legal domain. Our work emphasizes the need for syllogism-based reasoning to ensure logical consistency and adherence to legal principles.

**Transparency and Explainability.** Legal reasoning must be transparent and interpretable, particularly when used in sensitive or high-stakes domains. The metrics proposed in CitaLaw, including syllogism-based evaluation, aim to improve explainability by breaking down the reasoning process into logical components.

**Responsibility in System Deployment.** CitaLaw is intended as a research benchmark and should not be directly deployed in high-stakes legal decision-making without human oversight. While the benchmark aims to enhance the trustworthiness of LLM-generated responses, legal professionals should always verify the citations and legal interpretations provided by such systems. Misuse of automated systems without adequate validation could lead to inaccurate legal advice or unintended consequences in legal proceedings.

## 10 Acknowledgements

This work was funded by the National Key R&D Program of China (2023YFA1008704), the National Natural Science Foundation of China (62472426). Supported by fund for building world-class universities (disciplines) of Renmin University of China. Work partially done at Beijing Key Laboratory of Research on Large Models and Intelligent Governance, and Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE. Supported by the Beijing Social Science Foundation Planning Project (Grant No. 24GLC041), the Fundamental Research Funds for the Central Universities in UIBE (Grant No. 24QN06, 24PYTS22).

## References

Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. 2023. Exploring the state of the art in legal qa systems. *Journal of Big Data*, 10(1):127.

AI@Meta. 2024. Llama 3 model card.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 22–31.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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*.

Haolin Deng, Chang Wang, Xin Li, Dezhang Yuan, Junlang Zhan, Tianhua Zhou, Jin Ma, Jun Gao, and Ruifeng Xu. 2024. Webcites: Attributed query-focused summarization on chinese web search results with citations. *arXiv preprint arXiv:2403.01774*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023a. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *arXiv preprint arXiv:2311.03731*.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Preprint*, arXiv:2409.20288.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023. Xiezhi: Chinese law large language model. https://github.com/LiuHC0428/LAW_GPT.

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

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021b. 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.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. Expertqa: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

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. 2023a. fuzi.mingcha. https://github.com/irlab-sdu/fuzi.mingcha.

Yiquan Wu, Yuhang Liu, Yifei Liu, Ang Li, Siying Zhou, and Kun Kuang. wisdominterrogatory. Available at GitHub.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023b. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu

Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022a. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

Weijie Yu, Zhongxiang Sun, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen. 2022b. Explainable legal case matching via inverse optimal transport-based rationale extraction. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 657–668.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *Preprint*, arXiv:2309.11325.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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*, volume 34, pages 9701–9708.

Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Open chinese language pre-trained model zoo. Technical report.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *Preprint*, arXiv:2406.04614.

## A The Used Prompts

Figure 5 illustrates the prompts used in this paper, including $p_1$, $p_2$, $p_3$ in Eq. 1, Eq. 2 and Eq. 3.

## B More Details of Evaluated Models and Datasets

For the Legal LLMs, we choose (1) **fuzi.mingcha** (6B) (Wu et al., 2023a): It leverages unsupervised judicial corpora for training and uses syllogistic reasoning judgment data for fine-tuning. (2) **LexiLaw**[7] (6B): It specifically utilizes legal articles and legal reference books for training. (3) **Tailing**[8] (7B): It uses judicial text validation data, information extraction data, and judgment data for training. (4) **DISC-LawLLM** (13B) (Yue et al., 2023): In addition to fine-tuning with pairs, it also uses triplet data for fine-tuning to enhance the model's ability to leverage external knowledge. (5) **zhihai** (7B) (Wu et al.): It utilizes ChatGPT to modify the existing dataset and then performs secondary pre-training. (6) **LawGPT_zh** (6B) (Liu et al., 2023): It primarily uses scenario-based dialogues and knowledge-based question-answering data for fine-tuning based on LoRA. (7) **HanFei** (7B) (He et al., 2023): It is the first fully parameter-trained legal LLM in China. Because in the main experiment, CGG has the best overall performance, for the legal LLMs, we generate responses using CGG.

Table 4 and Table 5 are the website URLs and corresponding licenses of the evaluated models and datasets.

## C More Details on Implementation

Considering the length of legal texts and the input window for the LLMs is limited, all experiments in this paper are conducted using a zero-shot setting. We use the Chinese-performing-well Qwen2-1.5B (Yang et al., 2024)[9] to complete the MAUVE calculations. For RGUGE, We use version 1.0.1 of ROUGE for calculation. For BERTScore, we use bert-base-chinese (Devlin, 2018)[10] to compute it. Regarding sentence-BERT, we employ paraphrase-multilingual-MiniLM-L12-v2 (Reimers, 2019)[11].

[7] https://github.com/CSHaitao/LexiLaw

[8] https://github.com/DUTIR-LegalIntelligence/Tailing

[9] https://huggingface.co/Qwen/Qwen2-1.5B

[10] https://huggingface.co/google-bert/bert-base-chinese

[11] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

## D Human Evaluation

We hired four legal annotators from a Chinese university, all of whom have legal education backgrounds and are familiar with the cases in the dataset they need to annotate. We explained to the annotators that the data they annotated would be used for scientific research and paid them a reasonable remuneration based on local conditions. They are all graduate students from the judicial field, with practical experience in the legal profession. Two are male, two are female, aged between 24 and 30, and all have over five years of judicial theory study. Two annotators were responsible for the first stage of annotation, while the other two were responsible for the second stage, with all working together on the annotation process.

Table 6 shows a detailed description of each level used to evaluate the agreement of the NLI model with human evaluations.

## E Different Retrieval Models

Tables 7 and 8 present the performance of different retrieval models—Criminal-BERT, Civil-BERT, and BGE—on each metric for the CGG method across the two datasets. It can be observed that when Llama3 and Qwen2 are used as LLMs, BGE achieves the best performance as the retrieval model. Comparing the two datasets, on the Layperson dataset, where the questions are more general, Criminal-BERT and Civil-BERT, which focus on legal cases, perform relatively poorly. In contrast, on the Practitioner dataset, despite no structural or training improvements, Criminal-BERT and Civil-BERT achieve results comparable to BGE, highlighting the importance of legal knowledge in judicial QA tasks.

The differences between the two datasets also underscore the significance of selecting an appropriate retrieval model.

## F Different NLI Models

Figures 6 (a) and (b) show the entailment scores given by four legal LLMs as NLI models under different methods (CloseBook, CGG, ARG-Q, ARG-QA) and metrics($\text{Cita}_{\text{Law}}$, $\text{Cita}_{\text{S}}$, $\text{Cita}_{\text{B}}$, and $\text{Cita}_{\text{C}}$) when Qwen is used as the LLM. Similar conclusions to those in Section 6.5 can be drawn.

Answer the question based on the provided law article and cite it appropriately.
Only output the answer and citations, without including any additional content.
When citing the law article, use [A1] at the end of the relevant sentence.

Below is the provided law article:
Law article [A1]: **{Law article 1}**

Question: **{Question}**
Answer:

**(a) Layperson**

Answer the question based on the provided documents and cite them appropriately.
Only output the answer and citations, without including any additional content.
When citing precedent cases, use [C1], [C2], or [C3] at the end of the sentence. When citing the law article, use [A1].

Below are the provided documents:
Precedent case [C1]: **{Precedent case 1}**
Precedent case [C2]: **{Precedent case 2}**
Precedent case [C3]: **{Precedent case 3}**
Law article [A1]: **{Law article 1}**

Question: **{Question}**
Answer:

**(b) Practitioner**

Refine the text based on the references and only output the refined text.

Text: **{Answer}**
Reference: **{References}**

Refined text (only output the Refined text, without any additional content):

**(c) Response Refinement**

Answer the question.
Only output the answer without any additional content.

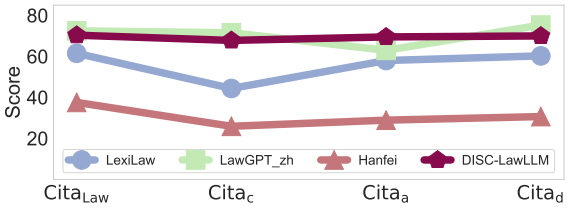
Question: **{Question}**
Answer:

**(d) Without Reference**

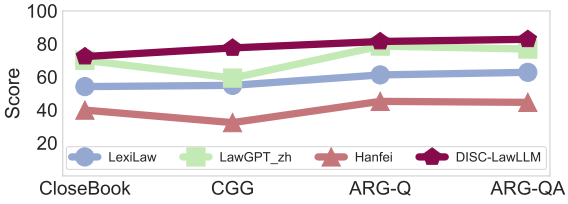
Figure 5: Prompts used in this paper. (a) The prompt $p_1$ is used to retrieve one law article in the Layperson dataset. (b) The prompt $p_1$ is used to retrieve one law article and three precedent cases in the Practitioner dataset. (c) The prompt $p_3$ is used to refine the LLM's answer based on references. (d) The prompt $p_2$ is used for LLM responses without references.

| Type | LLM | URL | Licence |
|---|---|---|---|
| Open domain | Qwen2-7B-Instruct | https://huggingface.co/Qwen/Qwen2-7B-Instruct | Apache-2.0 license |
| | Llam3-8B-Instruct | https://github.com/meta-llama/llama3 | META LLAMA 3 COMMUNITY License |
| Legal Domain | fuzi.mingcha | https://github.com/irlab-sdu/fuzi.mingcha | Apache-2.0 license |
| | DISC-LawLLM | https://github.com/FudanDISC/DISC-LawLLM | Apache-2.0 license |
| | LawGPT_zh | https://github.com/LiuHC0428/LAW-GPT | |
| | Hanfei | https://github.com/siat-nlp/HanFei | Apache-2.0 license |
| | Tailing | https://github.com/DUTIR-LegalIntelligence/Tailing | |
| | LexiLaw | https://github.com/CSHaitao/LexiLaw | MIT license |
| | zhihai | https://github.com/zhihaiLLM/wisdomInterrogatory | Apache-2.0 license |

Table 4: The LLM source URLs and licenses used by CitaLaw. The parts where the license is listed as empty indicate that the author has not provided a License.




(a) Metrics for CGG method with Layperson dataset.




(b) Methods for $\text{Cita}_{\text{Law}}$ metric with Practitioner dataset.

Figure 6: The performance of different NLI models when the LLM is Qwen.

| Type | Dataset | URL | Licence |
|---|---|---|---|
| Question | Layperson | https://github.com/open-compass/LawBench | Apache-2.0 license |
| | Practitioner | https://github.com/CSHaitao/LexEval | MIT License |
| Corpus | LeCaRD | https://github.com/myx666/LeCaRD | MIT License |
| | ELAM | https://github.com/ruc-wjyu/IOT-Match | MIT License |
| | CAIL2021-sfzy | https://github.com/china-ai-law-challenge/CAIL2021 | |
| | LJP-MSJudg | https://github.com/mly-nlp/LJP-MSJudge | |
| | fuzi.mingcha | https://github.com/irlab-sdu/fuzi.mingcha | Apache-2.0 license |
| | DISC-LawLLM | https://github.com/FudanDISC/DISC-LawLLM | Apache-2.0 license |
| | LawGPT_zh | https://github.com/LiuHC0428/LAW-GPT | |
| | Hanfei | https://github.com/siat-nlp/HanFei | Apache-2.0 license |

Table 5: The dataset source URLs and licenses used by CitaLaw. The parts where the license is listed as empty indicate that the author has not provided a License.

| **Score** | **Description** |
|---|---|
| **1** | No Entailment: The former does not entail the latter at all, with no logical connection between the two. |
| **2** | Weak Entailment: A partial entailment where the former somewhat relates to the latter, but the connection is weak and not fully conclusive. |
| **3** | Moderate Entailment: A moderate degree of entailment, meaning the former generally leads to the latter in most cases, but exceptions exist. |
| **4** | Strong Entailment: A strong logical relationship where the former can derive the latter in the vast majority of cases. |
| **5** | Complete Entailment: The former fully entails the latter in all cases, with an unambiguous and definitive logical connection between them. |

Table 6: Scoring Criteria for Human Evaluation of Entailment.

| Metric | | Fluency | Correctness | | | | | | | Citation | All |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Retriever | Mauve | Rouge-1 | Rouge-2 | Rouge-L | BERT-F | $\text{Correct}_c$ | $\text{Correct}_a$ | $\text{Correct}_d$ | $\text{Cita}_{Law}$ | **Avg** |
| Llama3 (Llam3-8B-Instruct) | Criminal | 37.44 | 18.07 | 2.18 | 13.15 | 61.71 | 64.03 | 63.56 | 64.36 | 80.34 | 44.98 |
| | Civil | 56.16 | 18.27 | 2.34 | 13.44 | 61.90 | 63.22 | 63.89 | 63.35 | 80.97 | 47.06 |
| | BGE | 61.01 | 23.97 | 6.05 | 17.91 | 65.94 | 67.29 | 77.31 | 74.95 | 86.70 | **53.46** |
| Qwen2 (Qwen2-7B-Instruct) | Criminal | 55.26 | 21.09 | 4.53 | 14.32 | 64.73 | 63.10 | 64.89 | 65.85 | 61.60 | 46.15 |
| | Civil | 52.44 | 20.48 | 4.16 | 13.81 | 64.45 | 61.79 | 64.94 | 65.62 | 59.88 | 45.29 |
| | BGE | 75.10 | 22.26 | 4.77 | 15.41 | 65.28 | 67.50 | 78.62 | 77.82 | 77.59 | **53.82** |

Table 7: Performance comparisons on retrieval models in the Layperson dataset when the method is CGG. The best performance is indicated in bold.

| Metric | | Fluency | Correctness | | | | | | | Citation | | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Retriever | Mauve | Rouge-1 | Rouge-2 | Rouge-L | BERT-F | $\text{Correct}_c$ | $\text{Correct}_a$ | $\text{Correct}_d$ | $\text{Cita}_{Law}$ | $\text{Cita}_c$ | $\text{Cita}_a$ | $\text{Cita}_d$ | **Avg** |
| Llama3 (Llam3-8B-Instruct) | Criminal | 34.25 | 25.79 | 7.86 | 19.42 | 65.03 | 66.27 | 76.30 | 76.82 | 70.59 | 66.41 | 70.09 | 69.47 | 54.03 |
| | Civil | 39.84 | 26.39 | 8.07 | 20.02 | 65.27 | 65.41 | 75.78 | 75.73 | 69.21 | 67.52 | 69.54 | 69.16 | 54.33 |
| | BGE | 36.37 | 26.15 | 7.84 | 19.55 | 65.60 | 67.19 | 76.36 | 77.73 | 73.58 | 68.23 | 67.87 | 67.65 | **54.51** |
| Qwen2 (Qwen2-7B-Instruct) | Criminal | 32.49 | 31.79 | 11.09 | 23.93 | 69.79 | 72.00 | 80.81 | 81.53 | 68.42 | 68.42 | 71.86 | 71.54 | 56.97 |
| | Civil | 33.37 | 31.67 | 11.06 | 23.84 | 69.63 | 73.35 | 80.57 | 81.27 | 69.11 | 66.41 | 70.09 | 69.47 | 56.65 |
| | BGE | 39.66 | 31.01 | 10.75 | 23.43 | 69.06 | 73.49 | 80.11 | 81.11 | 70.37 | 67.82 | 69.53 | 70.01 | **57.20** |

Table 8: Performance comparisons on retrieval models in the Practitioner dataset when the method is CGG. The best performance is indicated in bold.