Evaluating Test-Time Scaling LLMs for Legal Reasoning: OpenAI o1, DeepSeek-R1, and Beyond

Yinghao Hu [1*], Yaoyao Yu [2,4*], Leilei Gan [3†], Bin Wei [2,4†], Kun Kuang [1], Fei Wu [1,5]
[1]College of Computer Science and Technology, Zhejiang University
[2]Guanghua Law School, Zhejiang University
[3]School of Software Technology, Zhejiang University
[4]Law & AI Lab, Zhejiang University
[5]Shanghai AI Laboratory
{huyinghao, yaoyaoyu, leileigan, binwei, kunkuang, wufei}@zju.edu.cn

Abstract

Recent advances in test-time scaling of large language models (LLMs), exemplified by DeepSeek-R1 and OpenAI's o1, show that extending the chain of thought during inference can significantly improve general reasoning performance. However, the impact of this paradigm on legal reasoning remains insufficiently explored. To address this gap, we present the first systematic evaluation of 12 LLMs, including both reasoning-focused and general-purpose models, across 17 Chinese and English legal tasks spanning statutory and case-law traditions. In addition, we curate a bilingual chain-of-thought dataset for legal reasoning through distillation from DeepSeek-R1 and develop Legal-R1, an open-source model specialized for the legal domain. Experimental results show that Legal-R1 delivers competitive performance across diverse tasks. DeepSeek-R1 exhibits clear advantages in Chinese legal reasoning, while OpenAI's o1 achieves comparable results on English tasks. We further conduct a detailed error analysis, which reveals recurring issues such as outdated legal knowledge, limited capacity for legal interpretation, and susceptibility to factual hallucinations. These findings delineate the main obstacles confronting legal-domain LLMs and suggest promising directions for future research. We release the dataset and model at https://github.com/YinghaoHu/Legal-R1-14B.

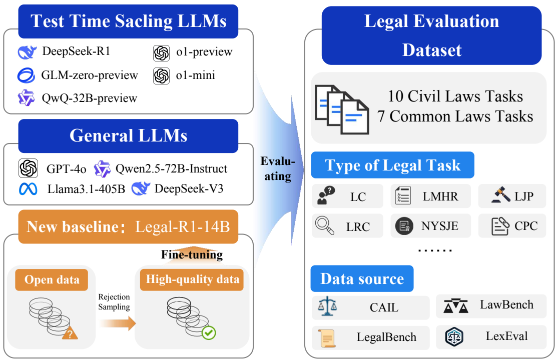


Figure 1: Overview of Work. The figure presents the 12 evaluated models together with representative task types and their data sources.

1 Introduction

Large language models (LLMs) have recently achieved near-human performance on an increasingly diverse set of benchmarks and application domains (Meta, 2024; Team, 2024; Openai, 2024a; team, 2025; Anthropic, 2025).

Across several flagship LLM model families, dedicated reasoning variants, such as OpenAI's o1 (Openai, 2024b) and DeepSeek-R1 (DeepSeek-AI, 2025) incorporate an explicit internal deliberation phase before producing a final answer. Fundamentally, these models extend the chain-of-thought (CoT) generated at inference time, thereby allocating increased computational resources per query.

The recent open-sourcing of DeepSeek-R1 further establishes an end-to-end paradigm for training reasoning-centric LLMs. Specifically, DeepSeek-AI (2025) proposes a four-stage pipeline: (i) cold-start pretraining, (ii) reasoning-oriented reinforcement learning (RL), (iii) rejection sampling-based supervised fine-tuning, and (iv) scenario-wide RL. This blueprint has inspired a new wave of test-time computation-intensive models, including QWQ-32B-Preview (Qwen Team, 2024) and GLM-zero-preview [1], which similarly extend reasoning traces, trading off computational cost for improved inference accuracy.

Contemporaneous work explores inference-time search strategies and training signals, such as Process Reward Models (Lightman et al., 2023; Uesato et al., 2022; Wang et al., 2024), self-corrective

[1] Equal contribution.

[2] Corresponding authors.

[1] https://bigmodel.cn/dev/api/normal-model/glm-zero-preview

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13759–13781
November 4-9, 2025 ©2025 Association for Computational Linguistics

RL schemes (Kumar et al., 2024), and Monte Carlo Tree Search (MCTS) and beam search variants (Feng et al., 2024; Trinh et al., 2024). While these approaches have not yet matched the reported performance of o1 (Openai, 2024b) and DeepSeek-R1, they nonetheless offer valuable insights for advancing the capabilities of reasoning-focused LLMs.

While the reasoning capabilities of LLMs have improved substantially in recent years, it would be premature to assume that such progress necessarily translates into strong performance on legal tasks. Legal reasoning imposes two simultaneous and demanding requirements: (i) the accurate synthesis of relevant statutes and case knowledge, and (ii) the rigorous application of this knowledge to novel and often complex fact patterns. Consequently, it remains uncertain whether models that perform well on general-purpose reasoning benchmarks can satisfy the domain-specific demands of legal reasoning. Although prior work has examined GPT-4 on specific legal tasks—such as legal text annotation (Savelka and Ashley, 2023), explaining legislative terminology (Savelka et al., 2023), and thematic analysis in empirical legal studies (Drápal et al., 2023)—these studies target narrow applications and assess models not purpose-built for reasoning. Consequently, a systematic evaluation of LLMs' legal reasoning across tasks spanning both statutory and case-law systems is still lacking.

To address this gap, we (i) present the first systematic evaluation of 17 legal reasoning tasks — seven in English and ten in Chinese — covering both test-time scaled and general-purpose LLMs; and (ii) construct a bilingual legal reasoning dataset using rejection sampling. Using this dataset, we progressively fine-tune DeepSeek-R1-Distill-Qwen-14B via supervised learning, resulting in Legal-R1, a domain-specific model with enhanced performance on legal tasks. Finally, we analyze errors across representative Chinese- and English-language tasks, identifying key challenges and future directions for improving legal reasoning in LLMs.

Our contributions can be summarized as follows:

1. Among the evaluated models, DeepSeek-R1 demonstrates superior performance in Chinese legal reasoning tasks. In English settings, both models perform similarly, achieving top results across several tasks. Nevertheless, even the strongest models continue to struggle with advanced reasoning tasks, such as those involving judicial ethics and complex tax calculations.

2. We introduce Legal-R1, developed through a progressive supervised fine-tuning strategy. It outperforms baseline models on the majority of Chinese and English legal tasks and exceeds DeepSeek-R1 on key tasks such as LC and IAPE, establishing a new standard for legal reasoning.

3. Our error analysis on representative Chinese and English legal tasks reveals key weaknesses, including outdated knowledge, limited legal understanding, and factual hallucinations. These results point to important directions for enhancing legal reasoning in LLMs.

2 Related Work

2.1 Legal Reasoning Benchmarks

Understanding the capabilities of LLMs in legal tasks, particularly legal reasoning, is a key focus of research (Blair-Stanek et al., 2023; Trozze et al., 2024), especially in tasks such as legal document generation (Iu and Wong, 2023), question answering (Hu et al., 2025), and judgment prediction (Gan et al., 2021, 2022; Jiang and Yang, 2023; Wei et al., 2025; Yuan et al., 2024, 2026). To facilitate legal reasoning evaluation, researchers have developed a diverse range of legal benchmarks, including LAR-ECHR (Chlapanis et al., 2024) and IL-TUR (Joshi et al., 2024). In addition, comprehensive benchmark suites such as LegalBench (Guha et al., 2023) for common-law tasks, LawBench (Fei et al., 2024) for civil-law evaluation, LexEval (Li et al., 2024a) for Chinese legal texts with ethical considerations, and Laiw (Dai et al., 2025), which emphasizes practice-oriented criteria, have been introduced.

However, legal systems differ across jurisdictions. Therefore, we construct a set of legal reasoning datasets covering both Chinese and U.S. legal systems to comprehensively evaluate the legal reasoning capabilities of current LLMs.

2.2 Test-Time Scaling

TTS has emerged as a powerful technique to boost the reasoning capabilities of LLMs during inference, without altering their underlying parameters or architecture. This paradigm has been adopted

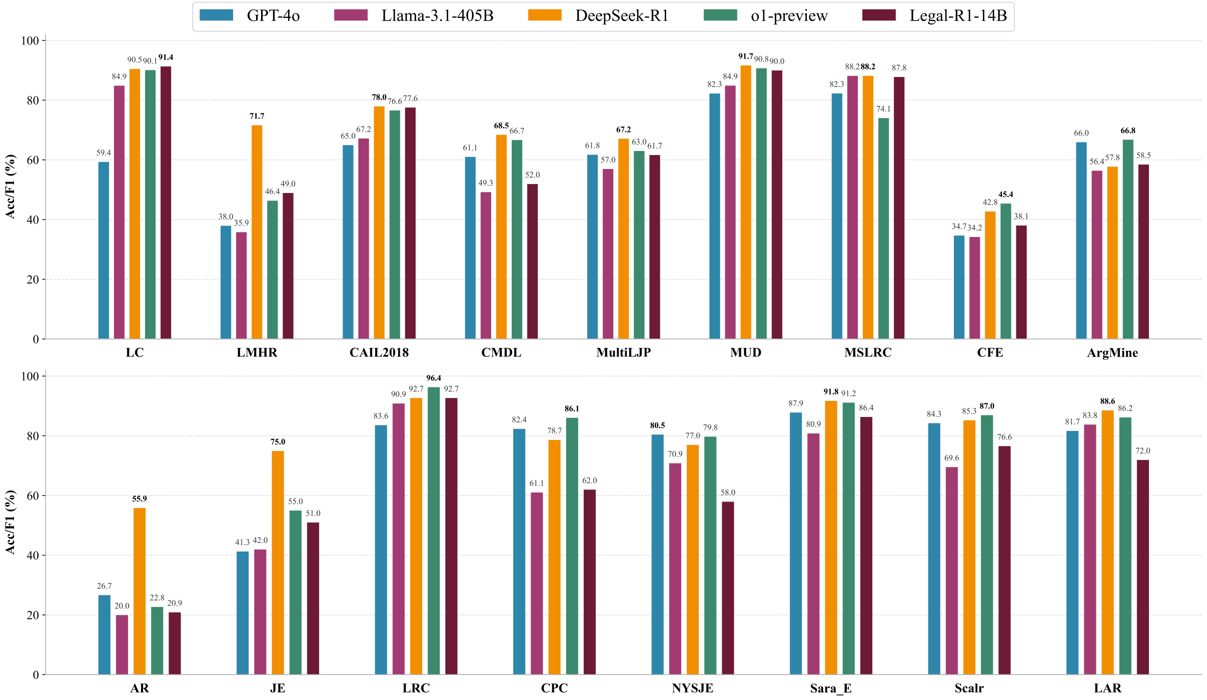


Figure 2: Overall Performance of LLMs on Chinese and English Legal Tasks. The figure shows the performance of representative LLMs on Chinese and English legal tasks. Inference models such as DeepSeek-R1 and o1-preview outperform traditional LLMs, while our model Legal-R1 achieves competitive performance.

by several prominent models, including OpenAI's o1 series (Openai, 2024b), Alibaba's QwQ-32B-Preview (Qwen Team, 2024), Zhipu AI's GLM-zero-preview, and DeepSeek-R1 (DeepSeek-AI, 2025). Several methods have been proposed to enable LLMs to leverage test-time scaling for enhanced reasoning. Verifier optimization, for instance, through process reward models (Lightman et al., 2023; Uesato et al., 2022; Wang et al., 2024), facilitates the incremental evaluation of reasoning steps, thereby boosting performance on complex tasks. Methods like STaR(Zelikman et al., 2022) and ReST (Singh et al., 2024) refine proposal distributions by fine-tuning models to generate more accurate answers without adding extra tokens. Self-critique techniques (Bai et al., 2022; Du et al., 2023; Madaan et al., 2023; Saunders et al., 2022) allow the model to iteratively refine its outputs. Search algorithms like Beam Search and Monte Carlo Tree Search(Feng et al., 2024; Trinh et al., 2024) further enhance exploration and solution accuracy. Despite their promise, the effectiveness of TTS-enhanced LLMs in legal tasks remains underexplored. This paper investigates whether their improved reasoning capabilities can transfer to the legal domain.

3 Evaluation Setting

3.1 Legal Reasoning Tasks

To comprehensively evaluate the legal reasoning capabilities of LLMs, we compile a benchmark comprising ten Chinese legal reasoning tasks rooted in the civil law tradition and seven English legal reasoning tasks based on the common law system.

Chinese tasks include Legal Calculation (LC), Legal Multi-hop Reasoning (LMHR), Legal Judgment Prediction (LJP), Multi-Defendant Legal Judgment Prediction (MDLJP), Multi-Defendant Charge Prediction (MDCP), Multi-segment Legal Reading Comprehension (MSLRC), Controversial Focus Extraction (CFE), Interactive Argument-Pair Extraction (IAPE), Article Recitation (AR), and Judicial Examination (JE).

English tasks include Legal Reasoning Causality (LRC), Citation Prediction Classification (CPC), NYS Judicial Ethics (NYSJE), Sara Numeric (Sara_N), Sara Entailment (Sara_E), Supreme Court Assessment of Legal Reasoning (Scalr), and Legal Argument Reasoning (LAR). Detailed descriptions of the datasets and tasks are provided in Appendix A.

Table 1: LLMs used for legal reasoning evaluation.

Category	Model	Source	Version
General LLMs	GPT-4o	OpenAI	2024-11
	Llama-3.1-405B	Meta	2024-07
	Qwen2.5-72B-Instruct	Alibaba	2024-09
	DeepSeek-V3	DeepSeek	2024-12
	Claude-Sonnet-4	Anthropic	2025-05
	Gemini 2.5 Pro	Google DeepMind	2025-06
Test Time Scaling LLMs	DeepSeek-R1	DeepSeek	2025-01
	OpenAI-o1-preview	OpenAI	2024-09
	OpenAI-o1-mini	OpenAI	2024-09
	GLM-zero-preview	Zhipu	2024-12
	QwQ-32B-Preview	Alibaba	2024-11
	DS.-R1-Distill-Qwen-14B	DeepSeek	2025-01
	Legal-R1-14B	Ours	2025-05

3.2 LLMs used for Evaluation

We evaluate LLMs from various providers across two categories: general-purpose models and models enhanced with test-time scaling. These models include both open- and closed-source implementations, cover diverse architectural designs such as dense and mixture-of-experts (MoE), and encompass both distilled and full-scale versions. A complete list is provided in Table 1.

4 Legal-R1

To transfer the reasoning capabilities of DeepSeek-R1 to the legal domain, we construct a high-quality legal reasoning dataset via rejection sampling guided by DeepSeek-R1. Based on this dataset, we fine-tune the DeepSeek-R1-Distill-Qwen-14B, yielding a domain-specific legal reasoning model, Legal-R1.

4.1 Reasoning Dataset Construction

4.1.1 Data Source

We collect the legal reasoning dataset covering both Chinese and U.S. legal contexts. For the Chinese law dataset, we curate a set of representative legal reasoning tasks, including legal calculation, legal multi-hop reasoning, interactive argument-pair extraction, legal judgment prediction, and multi-defendant legal judgment prediction. For the U.S. law dataset, we incorporate 143 tasks from LegalBench, excluding the English legal reasoning tasks listed in Table 9. The selected LegalBench tasks span six key categories of legal reasoning: (1) issue spotting, (2) rule recall, (3) rule application, (4) rule conclusion, (5) interpretation, and (6) rhetorical understanding. Together, these tasks comprehensively capture the essential dimensions of legal reasoning and provide a robust data foundation for adapting the model to both Chinese and U.S. legal domains.

4.1.2 Rejection Sampling

In the rejection sampling process, we first transform the legal reasoning dataset into a triple $P = (i, x, y)$, where i denotes the task description, x is the question, and y is the ground truth answer. For each question x, we use DeepSeek-R1 to generate multiple reasoning paths c and corresponding responses r according to the task description i. The generated response r is then compared with the ground truth y. If r matches y, the associated reasoning path c is retained. To control sampling cost, each question-answer pair is allowed up to three generation attempts. If none of the generated responses match the ground truth, the data is discarded. In cases where a match is found, the original triple is transformed into a quadruple $P = (i, x, c, y)$, where c represents the reasoning process. Finally, a total of 96,533 training samples, encompassing eight different tasks, are obtained using the aforementioned method.

This approach enables the construction of a high-quality legal reasoning dataset with faithful reasoning traces aligned to gold-standard answers, providing a strong foundation for training Legal-R1.

Regarding the potential bias that may arise from using DeepSeek-R1 to generate training data, we consistently follow the principle of minimizing bias and enhancing data quality. Detailed strategies are provided in Appendix D.1.

4.2 Training

During training, we employ a progressive supervised fine-tuning strategy based on DeepSeek-R1-Distill-Qwen-14B to obtain Legal-R1. In the first stage, the model is fine-tuned on a core legal reasoning task: legal judgment prediction. This task covers three key dimensions—charge prediction, relevant statute prediction, and sentence length prediction. Most downstream tasks, as well as many complex legal reasoning processes, are likely to rely on or relate closely to the capabilities established in legal judgment prediction. The completion of this process results in an intermediate model, denoted as M_{core}. Subsequently, M_{core} undergoes further fine-tuning on a comprehensive set of remaining legal tasks, integrating both Chinese and English legal reasoning datasets.

Table 2: Performance comparison of Chinese legal tasks under the closed-book setting. The best performance is highlighted in **bold**, while the second-best is underlined.

Model	LC↑	LMHR↑	CAIL2018↑	CMDL↑	MultiLJP↑	MUD↑	MSLRC↑	CFE↑	IAPE↑	AR↑	JE↑
General LLMs											
GPT-4o	59.40%	38.00%	65.00%	61.08%	61.79%	82.30%	82.33%	34.71%	65.99%	26.71%	41.33%
Llama3.1-405B	84.91%	35.86%	67.23%	49.26%	57.02%	84.94%	88.22%	34.25%	56.44%	20.03%	42.00%
Qwen2.5-72B-Instruct	80.34%	54.50%	76.50%	64.39%	61.41%	89.22%	84.97%	39.17%	60.50%	35.71%	50.67%
DeepSeek-V3	88.03%	45.00%	77.03%	63.94%	61.97%	87.67%	84.43%	40.00%	58.88%	36.05%	53.67%
Claude-Sonnet-4	89.86%	44.50%	77.48%	66.58%	60.04%	87.29%	87.19%	**48.74%**	60.00%	**63.58%**	53.24%
Gemini 2.5 Pro	91.03%	59.00%	**78.67%**	**69.35%**	**67.75%**	90.59%	87.88%	46.91%	**69.50%**	35.13%	65.76%
Test Time Scaling LLMs											
DeepSeek-R1	90.54%	**71.67%**	78.00%	68.48%	67.15%	**91.71%**	**88.23%**	42.80%	57.79%	55.91%	**75.00%**
o1-preview	90.13%	46.39%	76.63%	66.71%	63.05%	90.76%	74.07%	45.43%	66.83%	22.77%	55.03%
o1-mini	86.32%	27.00%	59.63%	42.59%	39.47%	84.02%	85.33%	39.78%	52.84%	12.62%	25.93%
GLM-zero-preview	72.22%	48.50%	71.98%	55.27%	55.82%	85.85%	78.56%	28.93%	57.00%	35.41%	48.67%
QwQ-32B-Preview	78.97%	56.00%	73.98%	60.70%	67.04%	87.04%	83.82%	27.75%	56.50%	34.93%	62.00%
DeepSeek-R1-Distill-Qwen-14B	91.03%	39.00%	72.03%	50.00%	56.50%	87.19%	87.03%	37.36%	57.00%	20.49%	48.67%
Legal-R1	**91.38%**	48.98%	77.60%	51.98%	61.70%	90.02%	87.85%	38.08%	58.50%	20.95%	51.05%

5 Experimentation

5.1 Experimental Setups

During training, we use 8 NVIDIA A800 GPUs. The learning rate is set to 1.0×10^{-5}, the cutoff length to 8,192, and bf16 precision is employed. The model is trained for 3 epochs.

For evaluation, LLMs' responses are retrieved via API requests. Tailored prompts are designed for each task to ensure a clear structure in the expected outputs. API request parameters are also adjusted according to the specific LLMs used. Detailed task prompts are provided in Appendix B.

5.2 Experimental Results

To provide an objective and fair assessment of the benefits that reasoning ability brings to legal tasks, we divide our evaluation into closed-book and open-book settings. In the closed-book setting, we provide no external knowledge or auxiliary information beyond the query, enabling a fairer measurement of a model's internal knowledge and overall reasoning ability. The open-book setting more closely reflects the workflow of practicing legal professionals and better decouples knowledge from reasoning.

We evaluate all tasks under the closed-book setting. Tables 2 and 5 present the performance of LLMs with and without TTS in Chinese and U.S. law, respectively. For knowledge-intensive tasks—such as LJP and MDLJP—we additionally evaluate under the open-book setting (Table 3).

5.2.1 Chinese Legal Task Results

As shown in Table 2, DeepSeek-R1 demonstrates consistently strong performance across a wide range of Chinese legal reasoning tasks. In particular, it excels in tasks that require logical inference and long-text comprehension, such as LMHR, MSLRC, and JE. On the other hand, our trained model(Legal-R1) shows consistent improvements over the baseline model (DeepSeek-R1-Distill-Qwen-14B), particularly on Chinese tasks. Notably, significant gains are observed on LMHR (+9.98%), CAIL2018 (+5.57%), and MDLJP (+5.2%). Although Legal-R1 still lags slightly behind DeepSeek-R1 in overall performance, it nevertheless demonstrates strong competitiveness on certain tasks (e.g., LC and IAPE). Considering that our model contains substantially fewer parameters, this highlights the effectiveness of our training strategy in specific scenarios.

Table 3: Performance comparison on knowledge-intensive tasks under Open-book and Closed-book settings. The left value corresponds to the Open-book setting, where the model is provided with gold-reference passages that directly contain the correct answer (ideal retrieval). The right value corresponds to the Closed-book setting.

Model	CAIL2018	CMDL	MultiLJP
DeepSeek-V3	77.13% / 77.03%	67.62% / 63.94%	70.69% / 61.97%
DeepSeek-R1	80.10% / 78.00%	72.21% / 68.48%	71.49% / 68.48%
GPT-4o	68.26% / 65.00%	64.10% / 61.08%	65.16% / 61.79%
Legal-R1	79.56% / 77.60%	59.76% / 51.98%	67.54% / 61.70%

As shown in Table 3, providing models with highly relevant external information improves accuracy on knowledge-intensive legal tasks. In Appendix D.2, we further explore the impact of retrieval quality on the final results.

Furthermore, we specifically examine the results of the **LJP subtask** to highlight model performance on this critical legal task. Based on Table 4, the variation in performance across the three subtasks — charge prediction, article prediction, and sentence prediction — highlights the models' differ-

Table 4: Evaluation results on the CAIL2018, CMDL, and MultiLJP datasets. Subscripts *cp*, *ap*, and *sp* denote Charge, Article, and Sentence Prediction.

Model	Task	Scores		
		F1$_{cp}$	**F1**$_{ap}$	**Acc**$_{sp}$
GPT-4o	CAIL2018	90.67	77.44	37.33
	CMDL	85.57	81.38	27.50
	MultiLJP	84.72	90.45	23.10
Llama3.1-405B	CAIL2018	86.00	83.00	41.33
	CMDL	77.86	58.47	20.91
	MultiLJP	74.13	81.40	25.90
DeepSeek-R1	CAIL2018	95.00	95.67	52.00
	CMDL	92.32	91.19	33.58
	MultiLJP	85.46	92.15	34.65
o1-preview	CAIL2018	94.33	96.67	48.33
	CMDL	91.00	91.29	30.07
	MultiLJP	84.11	89.14	27.68
Legal-R1	CAIL2018	94.33	96.33	51.00
	CMDL	85.59	76.84	8.12
	MultiLJP	83.91	88.68	24.80

ing capabilities in handling various forms of legal reasoning.

Charge prediction is generally a more straightforward task, often relying on explicit action verbs or key factual descriptions. As a relatively explicit task, it allows LLMs to make accurate predictions based on surface-level semantics and contextual cues. Consequently, both DeepSeek-R1 and Legal-R1 exhibit strong performance on this task, achieving F1 scores exceeding 80% across various datasets.

Article prediction is more challenging, as it requires the model not only to understand the act itself but also to match it with the appropriate legal provisions and their underlying logic. Since the same behavior may correspond to different legal articles depending on context, this task demands stronger analogical reasoning and structural comprehension from the model.

Sentence prediction does not have an absolute ground truth, as it depends on a range of subjective factors, including voluntary surrender, expressions of remorse, repeat offenses, and various mitigating or aggravating circumstances. As a hybrid task that combines elements of classification and regression, it poses greater challenges for LLMs, which often struggle with the nuanced judgments required for accurate sentencing estimation. As a result, both DeepSeek-R1 and Legal-R1 exhibit comparatively lower performance on this task.

5.2.2 English Legal Task Results

As shown in Table 5, LLMs generally perform better on English reasoning tasks than on Chinese ones. Among the models, the Test Time Scaling approach achieves superior results across most metrics, while DeepSeek-R1 delivers performance comparable to o1-preview. Certain tasks, such as LRC and Sara_E, appear relatively straightforward for LLMs. In contrast, the NYSJE task proves more challenging, with the highest observed accuracy reaching only 80.48%. Furthermore, most LLMs struggle on the Sara_N task, with DeepSeek-R1 standing out as a notable exception.

Our trained model shows overall improvements across the majority of tasks, except for LAR, where its performance slightly declines compared to the baseline. The improvements are more modest and consistently observed in English tasks than in Chinese ones. Specifically, we observe a 1.82% increase in performance on the LRC task and a 1.49% improvement on NYSJE. While our model generally underperforms compared to DeepSeek-R1 on most English tasks, it achieves results that are competitive with DeepSeek-R1 on the LRC task.

5.3 Error Analysis

To gain deeper insights into the limitations of DeepSeek-R1 and Legal-R1, we perform an error analysis on several representative tasks. For the Chinese tasks (IAPE, CFE, LJP, and AR), 30 error cases are randomly sampled from each task and analyzed by PhD students specializing in law. For the English tasks (CPC and NYSJE), all incorrect cases are examined by law PhD students to identify common error types. Examples of flawed reasoning processes are provided in Appendix C.

5.3.1 IAPE task

As shown in Figure 3, both DeepSeek-R1 and our proposed baseline model, Legal-R1, exhibit two primary error types in the IAPE task: Inconsistent Subjects and Indirect or Weak Rebuttals. Specifically, 93.0% of DeepSeek-R1's errors fall under Inconsistent Subjects and 7.0% under Indirect or Weak Rebuttals, whereas Legal-R1 shows 66.7% and 33.3% in these categories, respectively.

1. Inconsistent Subjects: This error arises when the subject chosen in the model's rebuttal is inconsistent with the subject presented in the plaintiff's argument. Such discrepancies often stem from the model's failure to grasp the core of the

Table 5: Performance comparison of English legal tasks under the closed-book setting. The best performance is highlighted in **bold**, while the second-best is underlined.

Model	LRC↑	CPC↑	NYSJE↑	Sara_N↓	Sara_E↑	Scalr↑	LAR↑
			General LLMs				
GPT-4o	83.64%	82.41%	**80.48%**	1.21	87.87%	84.30%	81.73%
Llama3.1-405B	90.91%	61.11%	70.89%	7.72	80.88%	69.59%	83.84%
Qwen2-72B-Instruct	87.27%	82.41%	70.89%	4.81	85.29%	77.19%	77.89%
DeepSeek-V3	90.91%	77.78%	75.00%	2.31	83.09%	77.19%	85.00%
Claude-Sonnet-4	87.45%	71.70%	64.04%	6.95	89.30%	80.12%	87.50%
Gemini 2.5 Pro	89.27%	84.26%	76.37%	5.98	88.97%	**89.47%**	87.00%
			Test Time Scaling LLMs				
DeepSeek-R1	92.73%	78.70%	77.05%	**0.25**	**91.79%**	85.28%	**88.60%**
o1-preview	**96.36%**	**86.11%**	79.79%	1.09	91.18%	86.98%	86.24%
o1-mini	87.27%	61.11%	66.78%	1.38	89.34%	73.53%	66.50%
GLM-zero-preview	83.64%	57.41%	65.41%	7.79	90.77%	70.76%	78.50%
QwQ-32B-Preview	78.18%	59.26%	64.73%	3.30	71.32%	73.41%	81.00%
DeepSeek-R1-Distill-Qwen-14B	90.91%	61.11%	56.51%	13.55	85.29%	75.44%	72.50%
Legal-R1	92.73%	62.04%	58.00%	12.50	86.40%	76.61%	72.00%

plaintiff's reasoning, frequently due to interference from complex legal background information.

2. Indirect or Weak Rebuttals: In these cases, although the model identifies the correct subject, the rebuttal produced is suboptimal, either because it lacks argumentative force or fails to directly engage with the core issues highlighted in the ground truth. This issue largely results from the model's inability to determine when to conclude its reasoning. As a consequence, it may over-extend the inference process and miss the critical point at which a direct and impactful response to the plaintiff's claim should occur, resorting instead to tangential or secondary arguments.

5.3.2 CFE Task

In the CFE task, we categorize errors into four levels based on the degree of deviation from the correct focus: complete deviation, major deviation, moderate deviation, and minor deviation. As illustrated in Figure 3, 67.0% of DeepSeek-R1's errors are complete deviations, 10.0% are major deviations, 10.0% are moderate deviations, and 13.0% are minor deviations. In comparison, Legal-R1 shows 60.0% complete deviations, 3.3% major deviations, 30.0% moderate deviations, and 6.7% minor deviations.

By analyzing the model's reasoning processes in these error cases, we identify a key underlying issue. Models that are not specifically trained in the legal domain often lack sufficient legal knowledge to accurately identify the core points of controversy. Although strong general-domain reasoning abilities lead to better performance in the CFE task compared to models with limited inference capabilities, they remain insufficient when the controversy involves specialized legal concepts such as duty of care or burden of proof.

5.3.3 LJP Task

In this task, we analyze the performance of sentence prediction across three datasets: CAIL2018, CMDL, and MultiLJP. Errors are categorized into two types: overestimation and underestimation of the predicted sentence length. By examining the reasoning processes of DeepSeek-R1 and Legal-R1, we identify the following primary causes of these errors:

1. Cumulative effects of hallucinations during reasoning: When the model makes an early misjudgment regarding factual details or legal applicability, subsequent steps tend to propagate this error. For instance, if a model incorrectly classifies an offense as "operating a casino" instead of "illegal gambling" due to flawed reasoning, this initial mistake may result in a substantially inaccurate sentence prediction.

2. Outdated or repealed legal provisions in training data: LLMs are typically trained on publicly available legal texts and internet sources. If the training data is not regularly updated, models may rely on outdated or invalid provisions, leading to erroneous predictions.

3. Overreliance on case similarity while overlooking critical differences: The models often analogize from previously encountered similar cases. While such analogical reasoning can be useful, it may lead to incorrect predictions when key factual or legal distinctions between the current case and prior examples are ignored.

5.3.4 AR Task

In the AR task, we identify four main types of errors: **article misidentification**, where the model substitutes content from one legal article for another; **content fabrication**, where the language model generates non-existent articles not present in the legal corpus; **omission of key provisions**, where essential parts of a legal article are left out; and **outdated references**, where the model cites outdated versions of legal articles that have since been amended or revised.

As illustrated in Figure 3, 73.0% of DeepSeek-R1's errors are article misidentifications, followed by 17.0% content fabrication, 7.0% omission of key provisions, and 3.0% outdated references. In contrast, Legal-R1 exhibits a different error distribution, with content fabrication accounting for the majority (70.0%), followed by 20.0% article misidentification, 6.7% omission of key provisions, and 3.3% outdated references.

For DeepSeek-R1, the high rate of article misidentification may stem from its multilingual training. Trained on both Chinese law and Anglo-American case law, it may confuse the two legal systems, leading to incorrect citations. For Legal-R1, the tendency to fabricate citations likely arises from its training data. Judicial documents usually include only the final legal provisions used by the court, not those considered and rejected. This causes the model to learn a rigid link between facts and a single statute. When it encounters new situations, it fills the gap by generating fabricated yet plausible laws.

5.3.5 NYSJE Task

As shown in Figure 3, **false positives** and **false negatives** account for nearly half of the incorrect cases. We further analyze the underlying causes.

We observe once again that factual hallucinations occur in the ethical guidelines generated by DeepSeek-R1. When lacking sufficient information to answer a question, DeepSeek-R1 tends to make unfounded assumptions — for example, adding contextual details that are not mentioned in the question. This behavior is neither rigorous nor reliable when it comes to answering legal questions.

For Legal-R1, most errors are attributable to the absence of task-specific information necessary for accurate responses. This may be due to limitations in the coverage of its domain-specific training data.

5.3.6 CPC Task

As shown in Figure 3, both DeepSeek-R1 and Legal-R1 exhibit confusion between "yes" and "no" responses in this task, without a pronounced bias toward either type of misclassification. We further analyze the reasons behind this:

1. Citation Factual Inaccuracies: We find that factual hallucinations about the content of citations occur during the reasoning process of LLMs. In addition, when the model lacks clarity about the details of the case, hallucinations may also arise, resulting in incorrect judgments.

2. Misunderstanding the Citation: In this task, correctly interpreting the citation is crucial for providing an accurate answer. Although LLMs have access to the full case details, any deviation in understanding the case can lead to an incorrect conclusion.

5.4 Ablation Study: Progressive SFT vs. Single-Stage SFT

We conduct an ablation study to compare the proposed Progressive SFT strategy with the traditional single-stage SFT approach using a dataset, ensuring a fair comparison by training both methods with equal total training steps and compute resources. Additionally, we evaluate a Base model without fine-tuning as a reference.

We present the results of the experiments on two sets of tasks: Chinese Legal Tasks (Table 6) and English Legal Tasks (Table 7).

Table 6: Chinese Legal Tasks Performance

Task	**Base**	**Single-Stage SFT**	**Progressive SFT**
LC	91.03%	91.34%	91.38% (+0.04%)
LMHR	39.00%	43.00%	48.98% (+5.98%)
CAIL2018	72.03%	77.40%	77.60% (+0.20%)
CMDL	50.00%	51.01%	51.98% (+0.97%)
MultiLJP	56.50%	61.52%	61.70% (+0.18%)
MUD	87.19%	84.40%	90.02% (+5.62%)
MSLRC	87.03%	87.14%	87.85% (+0.71%)
CFE	37.36%	37.83%	38.08% (+0.25%)
IAPE	57.00%	58.25%	58.50% (+0.25%)
AR	20.49%	20.81%	20.95% (+0.14%)
JE	48.67%	49.42%	51.05% (+1.63%)
Average	**58.75%**	**60.19%**	**61.64% (+1.45%)**

The results demonstrate the effectiveness of the Progressive SFT approach. For Chinese Legal Tasks, it achieves an average improvement of 1.45% over the single-stage approach and 2.89% over the base model. For English Legal Tasks, it outperforms the base model by 1.00% and the

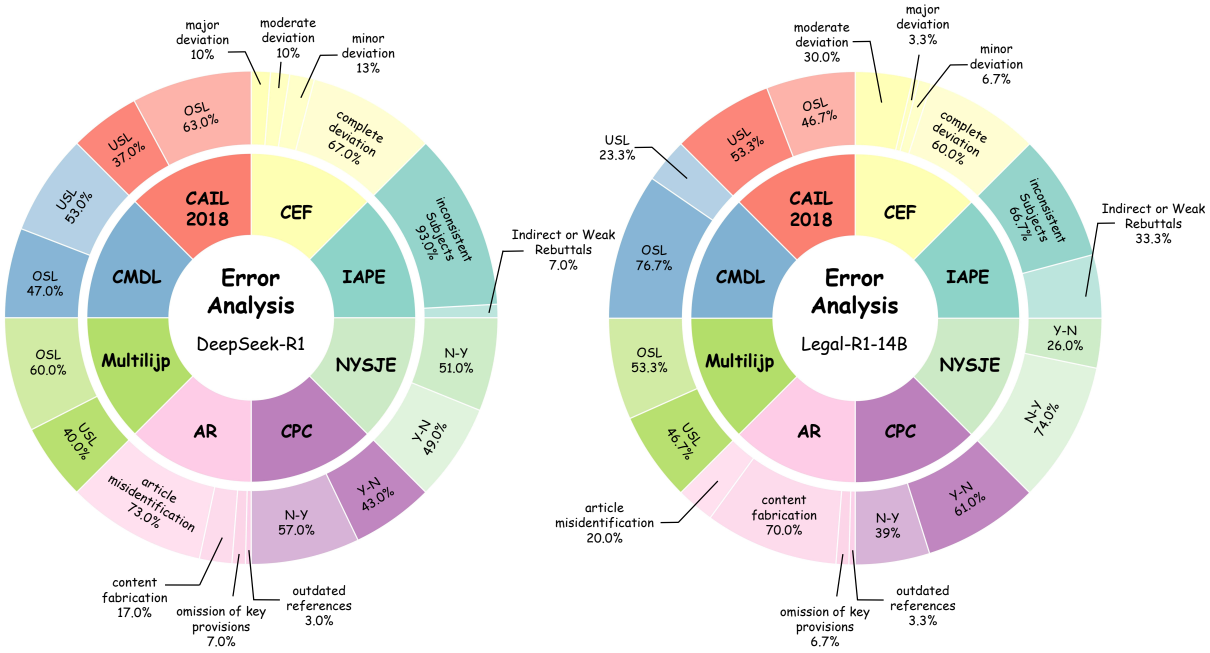


Figure 3: Error types across typical legal tasks.

Table 7: English Legal Tasks Performance

Task	Base	Single-Stage SFT	Progressive SFT
LRC	90.91%	94.55%	92.73% (-1.82%)
CPC	61.11%	58.35%	62.04% (+3.69%)
NYSJE	56.51%	56.85%	58.00% (+1.15%)
Sara_N	13.55	12.97	12.50 (-0.47)
Sara_E	85.29%	85.84%	86.40% (+0.56%)
Scalr	75.44%	76.42%	76.61% (+0.19%)
LAR	72.50%	71.50%	72.00% (+0.50%)
Average	**73.63%**	**73.92%**	**74.63% (+0.71%)**

single-stage approach by 0.71%. Although the gains on English tasks are smaller, the consistent improvements on most tasks highlight the robustness of the Progressive SFT strategy.

6 Conclusion

This study presents a comprehensive evaluation of 12 LLMs across 17 Chinese and English legal reasoning tasks and introduces Legal-R1, an open-source model tailored for legal reasoning. Our experiments confirm that test-time scaling improves overall reasoning performance. DeepSeek-R1 remains among the strongest on both Chinese and English tasks, while Legal-R1, trained on a curated legal-reasoning dataset, matches or surpasses test-time scaling models on several key tasks and establishes a competitive open-source baseline. Error analysis reveals persistent challenges shared by general-purpose and domain-specific models, including outdated or incomplete legal knowledge, misinterpretation of citations, and factual hallucinations. Expanding high-quality, up-to-date, multilingual chain-of-thought legal datasets, integrating retrieval or external knowledge bases for fact verification, and developing more robust reasoning architectures will be essential for improving the reliability and practicality of LLMs in legal reasoning.

Limitations

Although our benchmark encompasses a variety of legal reasoning tasks in both Chinese and English, it may not fully capture the breadth and complexity of legal reasoning encountered in real-world practice. Certain tasks, such as issue identification and ethical judgment, involve a degree of subjectivity, where even domain experts may differ in their evaluations. In such cases, existing automatic evaluation metrics may fall short of accurately reflecting the quality of legal reasoning in model outputs. Furthermore, while our baseline models achieve encouraging results, there remains substantial room for improvement. We believe future work can build on this foundation by broadening task coverage, developing more nuanced evaluation methodologies, and enhancing model performance in complex legal scenarios.

Ethics Statement

Given the sensitive nature of the legal domain, the application of artificial intelligence in this field requires rigorous ethical management. To address potential ethical concerns, we adopt the following measures. In particular, to prevent the leakage of private information (e.g., personal names), we anonymize or replace sensitive information with neutral third-person references when constructing both training datasets and evaluation benchmarks. This ensures that our research adheres to principles of privacy protection and responsible AI development.

Acknowledgements

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2025C02037).

References

Anthropic. 2025. Claude 3.7 sonnet and claude code. Accessed: 2025-02-25.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning?

Odysseas S. Chlapanis, Dimitrios Galanis, and Ion Androutsopoulos. 2024. LAR-ECHR: A new legal argument reasoning task and dataset for cases of the European court of human rights. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 267–279, Miami, FL, USA. Association for Computational Linguistics.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. LAiW: A Chinese legal large language models benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10738–10766, Abu Dhabi, UAE. Association for Computational Linguistics.

Sybren de Kinderen and Karolin Winter. 2024. Towards taming large language models with prompt templates for legal grl modeling. In *International Conference on Business Process Modeling, Development and Support*, pages 213–228. Springer.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Jakub Drápal, Hannes Westermann, and Jaromir Savelka. 2023. Using large language models to support thematic analysis in empirical legal studies.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2024. Alphazero-like tree-search can guide large language model decoding and training.

Leilei Gan, Kun Kuang, Yi Yang, and Fei Wu. 2021. Judgment prediction via injecting legal knowledge into neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12866–12874.

Leilei Gan, Baokui Li, Kun Kuang, Yating Zhang, Lei Wang, Luu Anh Tuan, Yi Yang, and Fei Wu. 2022. Exploiting contrastive learning and numerical evidence for confusing legal judgment prediction. *arXiv preprint arXiv:2211.08238*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yinghao Hu, Leilei Gan, Wenyi Xiao, Kun Kuang, and Fei Wu. 2025. Fine-tuning large language models for improving factuality in legal question answering. *arXiv preprint arXiv:2501.06521*.

Wanhong Huang, Yi Feng, Chuanyi Li, Honghan Wu, Jidong Ge, and Vincent Ng. 2024. CMDL: A large-scale Chinese multi-defendant legal judgment prediction dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5895–5906, Bangkok, Thailand. Association for Computational Linguistics.

Kwansai Iu and Vanessa Man-Yi Wong. 2023. Chatgpt by openai: The end of litigation lawyers? *SSRN Electronic Journal*.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 417–421, New York, NY, USA. Association for Computing Machinery.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. IL-TUR: Benchmark for Indian legal text understanding and reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11460–11499, Bangkok, Thailand. Association for Computational Linguistics.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. 2024. Training language models to self-correct via reinforcement learning.

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

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step.

Yougang Lyu, Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang, and Zhaochun Ren. 2023. Multi-defendant legal judgment prediction via hierarchical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2198–2209, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Meta. 2024. Introducing llama 3.1: Our most capable models to date. Accessed: 2025-03-01.

Openai. 2024a. Hello gpt-4o. Accessed: 2025-03-01.

Openai. 2024b. Learning to reason with llms. Accessed: 2025-03-01.

Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown. Blog post. Accessed: 2025-03-01.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators.

Jaromir Savelka and Kevin D. Ashley. 2023. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6.

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

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. 2024. Beyond human data: Scaling self-training for problem-solving with language models.

Gemini team. 2025. Gemini 2.0 is now available to everyone. Accessed: 2025-03-01.

Qwen Team. 2024. Qwen2.5: A party of foundation models! Accessed: 2025-03-01.

Trieu Trinh, Yuhuai Tony Wu, Quoc Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625:476–482.

A. Trozze, T. Davies, and B. Kleinberg. 2024. Large language models in cryptocurrency securities cases: Can a GPT model meaningfully assist lawyers? *Artificial Intelligence and Law*.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Bin Wei, Yaoyao Yu, Leilei Gan, and Fei Wu. 2025. An llms-based neuro-symbolic legal judgment prediction framework for civil cases. *Artificial Intelligence and Law*, pages 1–35.

Xiao Wei, Qi Xu, Hang Yu, Qian Liu, and Erik Cambria. 2024. Through the MUD: A multi-defendant charge prediction benchmark with linked crime elements. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2864–2878, Bangkok, Thailand. Association for Computational Linguistics.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7577–7597.

Weikang Yuan, Kaisong Song, Zhuoren Jiang, Junjie Cao, Yujie Zhang, Chengyuan Liu, Jun Lin, Ji Zhang, Kun Kuang, and Xiaozhong Liu. 2026. A multi-agent framework with legal event logic graph for multi-defendant legal judgment prediction. *Information Processing & Management*, 63(1):104319.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning.

Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Jec-qa: A legal-domain question answering dataset.

A Appendix A

A.1 Chinese Legal Tasks

Legal Calculation: The legal calculation task involves answering multiple-choice questions that require legal computations. For each question, the model must select the single correct option from A, B, C, or D. This task is evaluated on the LC dataset derived from LexEval (Li et al., 2024b), a comprehensive Chinese legal benchmark for assessing LLMs, using accuracy as the evaluation metric.

Legal Multi-hop Reasoning: This task assesses the legal knowledge and reasoning capabilities of LLMs. The input consists of multiple-choice questions related to legal matters, and the model's output is the correct answer(s) from the provided options, which may include one or more correct choices. The LMHR dataset, sourced from LexEval, is used for this task, with accuracy as the evaluation metric.

Legal Judgment Prediction: This task focuses on legal judgment prediction for single-defendant cases based on the CAIL2018 dataset. The input includes a detailed description of case facts and defendant information, while the output provides judgment results for three subtasks: charge prediction, article prediction, and sentence prediction. The evaluation employs the same metrics as those used in CAIL2024 [2], with the calculations detailed as follows:

For a given case c with n defendants, consider a defendant d who is charged with m_1 crimes. If the model predicts m_2 crimes for this defendant, with m_3 of them being correct, the precision (P), recall (R), and F1 score ($F1$) for the charge and article prediction subtasks for this defendant are defined as follows:

$$P_d^c = \frac{m_3}{m_2}, R_d^c = \frac{m_3}{m_1}, F1_d^c = \frac{2 \cdot P_d^c \cdot R_d^c}{P_d^c + R_d^c} \quad (1)$$

The P, R, and F1 Score for this case are calculated as follows:

$$P_c = \frac{\sum_{i=1}^{n} P_i^c}{n} \quad (2)$$

$$R_c = \frac{\sum_{i=1}^{n} R_i^c}{n} \quad (3)$$

$$F1_c = \frac{\sum_{i=1}^{n} F1_i^c}{n} \quad (4)$$

For the entire dataset, these metrics are weighted by $w_c = \log_2 n$:

$$P = \frac{\sum w_c P_c}{\sum w_c}, R = \frac{\sum w_c R_c}{\sum w_c}, F1 = \frac{\sum w_c F1_c}{\sum w_c} \quad (5)$$

The metric of sentence prediction for case c is evaluated using the accuracy metric. For a given

[2] https://github.com/china-ai-law-challenge/CAIL2024/tree/main/drdz

Table 8: Chinese Legal Tasks

Task	Dataset	Source	Metric	Test Size
Legal Calculation(LC)	LC	LexEval	Acc	234
Legal Multi-hop Reasoning(LMHR)	LMHR	LexEval	Acc	200
Legal Judgment Prediction(LJP)	CAIL2018	CAIL2018	F1	300
Multi-Defendant Legal Judgment Prediction(MDLJP)	CMDL	Huang et al. (2024)	F1	300
Multi-Defendant Legal Judgment Prediction(MDLJP)	MultiLJP	Lyu et al. (2023)	F1	300
Multi-Defendant Charge Prediction(MDCP)	MUD	Wei et al. (2024)	F1	175
Multi-segment Legal Reading Comprehension(MSLRC)	MSLRC	CAIL2021	F1	200
Controversial Focus Extraction(CFE)	CFE	LAIC2021	F1	200
Interactive Argument-Pair Extraction(IAPE)	ArgMine	CAIL2023	Acc	200
Article Recitation(AR)	AR	LawBench	Rouge-L	200
Judicial Examination(JE)	JE	JEC-QA	Acc	300

case c with n defendants, if k defendants have correctly predicted sentences, then:

$$Acc_c = \frac{k}{n} \tag{6}$$

The sentence accuracy for the entire dataset is:

$$Acc = \frac{\sum w_c Acc_c}{\sum w_c}, \quad w_c = \log_2 n \tag{7}$$

Finally, the overall F1 score, combining the metrics for the three subtasks, is calculated as:

$$\text{F1} = 0.3 \times \text{F1}_{\text{cp}} + 0.3 \times \text{F1}_{\text{ap}} + 0.4 \times \text{Acc}_{\text{sp}} \tag{8}$$

Here, F1_{cp} and F1_{ap} denote the F1 scores for charge and article prediction, respectively, and Acc_{sp} represents the sentence prediction accuracy.

Multi-Defendant Legal Judgment Prediction: This task focuses on predicting legal judgments in cases involving multiple defendants. The task utilizes two datasets: CMDL from Huang et al. (2024) and MultiLJP from Lyu et al. (2023), and employs the same evaluation metrics as used in the LJP task.

Multi-Defendant Charge Prediction: This task focuses on predicting charges for multiple defendants. Given the case facts as input, the goal is to determine the charges committed by each defendant. The dataset used is MUD from Wei et al. (2024), and the evaluation metric is analogous to that of the charge prediction subtask in the LJP task.

Multi-segment Legal Reading Comprehension: This task involves multi-segment questions, where the answers are derived by extracting and combining multiple segments from the legal text. The dataset employed is MSLRC from CAIL2021. To evaluate LLMs performance on this task, we designed a metric tailored to its characteristics. Here, both the ground truth $G = \{g_1, g_2, \ldots, g_n\}$ and the model output $E = \{e_1, e_2, \ldots, e_m\}$ are lists of legal elements that answer the question within its legal context. A pre-trained language model is used to automatically assess the semantic similarity between the elements in G and E. Finally, the F1 score is computed as the evaluation metric, calculated as follows:

$$P = \frac{N}{m}, R = \frac{N}{n}, F1 = \frac{2PR}{P+R} \tag{9}$$

where N represents the number of correctly predicted legal elements in E, m is the total number of elements in the output E, and n is the total number of elements in the ground truth G.

Controversial Focus Extraction: This task entails identifying dispute issues based on the claims and defenses from both the plaintiff and defendant. The output is a list of controversial focus indices extracted from the case facts. LLMs performance is assessed using the F1 score, calculated similarly to the MSLRC task. However, rather than relying on a pre-trained language model for semantic interpretation, we directly verify whether the predicted indices match the ground truth indices.

Interactive Argument-Pair Extraction: This task aims to extract interaction argument pairs by identifying the defense counter-argument that corresponds to a given plaintiff's argument. The input comprises the plaintiff's argument along with five candidate defense arguments, and the output is the selected counter-argument. Performance is measured using accuracy.

Article Recitation: This task assesses LLMs' ability to recall legal knowledge by prompting them to recite the content of legal articles based on their reference numbers. It examines their proficiency in memorizing key legal concepts, terminology, and provisions. The dataset is sourced from the comprehensive LawBench evaluation benchmark

Fei et al. (2024), and Rouge-L is employed as the evaluation metric.

Judicial Examination: This task requires LLMs to output the final answers to the questions from JEC-QA (Zhong et al., 2019), which is the largest question answering dataset in the legal domain, collected from the National Judicial Examination of China. We randomly test the 300 cases from the concept comprehension questions and scenario analysis questions, which require the ability of logical reasoning. The performance of LLMs is measured using accuracy.

A.2 English Legal Tasks

The English legal reasoning tasks are mainly sourced from LegalBench (Guha et al., 2023), a collaboratively constructed legal reasoning benchmark consisting of 162 tasks covering six different types of legal reasoning. Besides, we also add a new legal argument reasoning task proposed by Chlapanis et al. (2024). The tasks are listed as follows:

Legal Reasoning Causality: This task aims to classify whether an excerpt from a district court opinion relies on statistical evidence in its reasoning.

Citation Prediction Classification: The task requires determining whether a given case citation supports a legal statement, based on the provided legal statement and citation.

NYS Judicial Ethics: In this task, LLMs are required to determine whether a question violates judicial ethics in the New York State Unified Court System. The dataset consists of real ethical scenarios, reformulated into questions to evaluate the models' understanding of ethical rules and their application in different judicial contexts.

Sara Numeric: In this task, the LLMs should determine how much tax an individual owes given a statute and accompanying facts. The dataset in this task is from the StAtutory Reasoning Assessment(SARA), it contains a set of statutes and summaries of facts paired with a numerical question. Additionally, we use Mean Squared Error (MSE) as the evaluation metric for this task. To reduce the impact of extreme values, we calculate the MSE after applying the logarithmic transformation (log1p) to the true and predicted values.

Sara Entailment: In this task, given a statute, a fact, and an assertion, LLMs are required to determine if the assertion is "entailed" by the fact and statute. The dataset in this task is also from SARA, which tests the ability to reason about summaries of facts and statutes, in the context of US federal tax law.

Legal Calculation

```
Please read the following multiple-choice question and provide the correct option without explaining the reason. Please only provide the letter of the answer (A, B, C, D).
{id}
{input}
Please strictly follow the format below to provide the prediction result in a JSON file! The format is as follows:
{{
  "id": "{id}",
 "answer": ""
}}
Chinese:
请阅读以下选择题给出正确选项，不要解释原因。请只给出答案的序号(A,B,C,D)。
{id}
{input}
请严格按下列格式给出预测结果，以json的格式输出结果文件！格式如下：
{{
  "id": "{id}",
 "answer": ""
}}
```

Figure 4: The prompt for LC dataset.

Supreme Court Assessment of Legal Reasoning: In this task, the model must select, from a set of candidates, the holding statement that best answers a specific legal question. Each question represents an issue reviewed in a particular Supreme Court case, and the model must identify the holding statement that most accurately addresses it. This task is designed to assess legal reasoning by emphasizing the understanding of legal language over rote memorization of legal knowledge.

Legal Argument Reasoning: This task involves selecting the appropriate subsequent statement from multiple choices within a sequence of legal arguments presented during Court proceedings, based on the case facts. The input consists of a case description, a specific argument related to the case, and several potential candidate arguments. The objective is to determine which candidate argument logically continues the given argument.

B Appendix B

In this section, we present the instructions provided to LLMs for evaluating legal tasks in both Chinese and English. For details, see Figures 4–23.

C Appendix C

This section presents examples of flawed reasoning processes observed in several representative tasks.

Table 9: English Legal Tasks

Task and Dataset	Source	Metric	Test Size
Legal Reasoning Causality(LRC)	LegalBench	Acc	55
Citation Prediction Classification(CPC)	LegalBench	Acc	108
NYS Judicial Ethics(NYSJE)	LegalBench	Acc	292
Sara Numeric(Sara_N)	LegalBench	Mse	96
Sara Entailment(Sara_E)	LegalBench	Acc	272
Supreme Court Assessment of Legal Reasoning(Scalr)	LegalBench	Acc	172
Legal Argument Reasoning(LAR)	Chlapanis et al. (2024)	Acc	200

Legal Multi-hop Reasoning

Please read the following multiple-choice question and provide the correct option(s) without explaining the reason. Please only provide the letter(s) of the answer (A, B, C, D).
Note: The correct answer(s) may include one or more options.
{id}
{input}
Please strictly follow the format below to provide the prediction result in a JSON file! The format is as follows:

```
{
"id": "{id}",
"answer": ""
}
```

Chinese:
请阅读以下选择题给出正确选项，不要解释原因。请只给出答案的序号(A,B,C,D)。
注意：正确答案可能具备一个或多个。
{id}
{input}
请严格按下列格式给出预测结果，以JSON的格式输出结果文件！格式如下：

```
{{
   "id": "{id}",
   "answer": ""
}}
```

Figure 5: The prompt for LMHR dataset.

C.1 IAPE task

Table 10 illustrates typical flawed reasoning identified in the IAPE task.

C.2 CFE task

Table 11 illustrates typical flawed reasoning identified in the CFE task.

C.3 LJP task

Table 12 illustrates typical flawed reasoning identified in the LJP task.

C.4 AR task

Table 13 illustrates typical flawed reasoning identified in the AR task.

C.5 CPC task

Table 14 illustrates typical flawed reasoning identified in the CPC task.

C.6 NYSJE task

Table 15 illustrates typical flawed reasoning identified in the NYSJE task.

Legal Judgment Prediction(CAIL2018)

##Assume you are a judge. Based on the provided charge, relevant legal provisions, and case details, make a judgment prediction for the given defendant.
The list of charges is as follows: [xx]
The list of relevant legal provisions is as follows: [xx]
##Here’s an example:
id: 307,
The case facts: The People's Procuratorate of Jining District, Ulanqab City, charges that at approximately 21:10 on February 28, 2018, the defendant Guo drove a white Dongfeng Nissan brand small ordinary passenger car on the road after consuming alcohol. When driving to the area of Xingfu Road in Jining District, Xingfu Square... The criminal facts are clear, and the evidence is substantial and sufficient; criminal responsibility should be pursued in accordance with the law. A sentence is now requested according to the law.
Defendant: Guo
Judgments:

```
{
   "id": 307,
   "judgments": {
      "charges": ["Dangerous Driving"],
      "articles": ["133"],
      "penalty": {
         "imprisonment": 1,
         "death_penalty": false,
         "life_imprisonment": false
      }
   }
}
```

In this case, "charges" indicates the crimes committed by the defendant, and "articles" indicates the legal provisions the defendant has violated. The "penalty" indicates the sentence given to the defendant, represented as key-value pairs. For example: {"imprisonment": [int], "death_penalty": [bool], "life_imprisonment": [bool]}. This represents the length of the prison term, whether the defendant will face the death penalty, or if they will receive a life sentence. The duration is in months, for example, one year of imprisonment would be represented as "imprisonment": 12.
##Based on the above example, please make a judgment prediction for the defendant in the following case. The defendant is only involved in one charge in the current case:
id: {id}
Case facts: {fact}
Defendant: {defendants}
Judgments:
Please strictly follow the format below to give the prediction result, and output it as a JSON file! The format example is as follows:

```
{
   "id": {id},
   "judgments": {
      "charges": ["xxx"],
      "articles": ["xxx"],
      "penalty": {
         "imprisonment": xxx,
         "death_penalty": xxx,
         "life_imprisonment": xxx
      }
   }
}
```

Figure 6: The prompt for CAIL2018 dataset.

Table 10: Examples of flawed reasoning in the IAPE task.

Type	Example
Missing Plaintiff's Core Argument	I need to identify which defense argument directly responds to or interacts with the plaintiff's compensation claim. The plaintiff is seeking joint compensation, so I need to look for any defense that directly addresses the liability for compensation or the amount. bc_5 Insurance Company has made it clear that it will assume responsibility within the scope of compulsory insurance, which directly corresponds to the compensation amount and falls within the scope of the insurance company's liability. (The core subject is compensation.)
Flawed and Excessive Reasoning	Dispute over Liability Allocation: The plaintiff asserts through evidence that Sha should bear full responsibility, whereas bc_2 proposes that Deng xx should bear secondary responsibility. If Deng xx's responsibility holds true, the scope of Sha's responsibility will inevitably be curtailed (e.g., primary responsibility rather than full responsibility), directly affecting whether he constitutes a traffic accident crime (since this crime typically requires the perpetrator to bear major or full responsibility). Confrontation on Conviction and Sentencing: The plaintiff demands a heavier punishment for Sha, while bc_2 argues that Deng xx should not be deemed criminal and should be given a lighter punishment based on circumstances such as surrender and compensation. If Deng xx's responsibility is acknowledged, it might undermine the foundation for Sha's conviction and subsequently weaken the rationality of imposing a more severe punishment. Therefore, bc_2 is the sole argument that forms a substantial interaction with the plaintiff. By introducing the responsibility sharing of another defendant, it directly challenges the plaintiff's core assertion regarding Sha's full responsibility and the establishment of criminal liability. (A lengthy overthinking reasoning process)

Table 11: Examples of flawed reasoning in the CFE task.

Type	Example
Lacking Sufficient Legal Knowledge	Additionally, the division of vehicles and real estate in property division may involve "real estate division" 118, but index 45 has already covered the overall property division. Moreover, the defendant's request for an equal division of common property also falls within the scope of property division. There may also be "determination of property nature" 44, but there is no mention of a dispute over the nature of the property in the case, only the issue of division.
Struggling to Understand Disputed Issues	The second part is the issue of advance payment for medical expenses, amounting to 7,900 yuan, which needs to be dealt with in this case. This involves the determination of medical expenses, which falls under "medical and nursing expense determination," corresponding to index 14. However, it might also involve the determination of compensation items or amounts. Looking more closely at the dictionary, there are "compensation item determination" (6) and "loss compensation amount determination" (7), but since medical expenses are specific medical and nursing costs, the more direct reference is index 14, "medical and nursing expense determination."
Confusing Disputed Issues	The core dispute in the case facts lies in the appellant's claim that the loan in question is a fictitious debt and believes that even if the loan is real, it belongs to PER's personal debt rather than the joint debt of the couple. The appellee, on the other hand, claims that the loan is real and valid. According to the dispute focus dictionary, "determination of joint debt of husband and wife" (index 54) directly corresponds to the dispute between the two parties over whether the debt belongs to the joint debt during the marriage relationship, which is the most fitting dispute focus. (The ground truth is Joint Debt Claim of Husband and Wife / Common Debt during the Existence of Marriage Relationship.)

D Appendix D

D.1 Strategies to Minimize Potential Bias and Improve Data Quality

To minimize potential bias and improve data quality, we adopted several strategies during data collection. First, we incorporated a domain-specific legal terminology dictionary to guide DeepSeek-R1 toward generating legally relevant responses, as prior studies have shown that such dictionaries can significantly enhance terminological accuracy

Table 12: Examples of flawed reasoning in the LJP task.

Type	Example
Confusing the Crimes	Based on provided charge categories, possible charges include contract fraud and fraud. The defendants obtained vehicles through a leasing agreement, then fabricated IOUs and sold the vehicles with the intent to illegally possess them, which meets the constituent elements of contract fraud. Contract fraud involves deceiving parties to obtain property during the signing or performance of contracts. Here, fraudulent acts were committed during the leasing process. Therefore, the charge should be contract fraud, corresponding to Article 224 of the Criminal Law.
Outdated Legal Provision	Next, determine the sentence. According to Article 303 of the Criminal Law, the crime of operating a gambling establishment generally carries a sentence of up to 3 years of imprisonment, criminal detention, or controlled release, plus fines; for severe cases, 3 to 10 years. In this case, their illegal gains exceed 50,000 yuan, constituting a "relatively large amount" but likely not qualifying as a "severe circumstance".
Based on Similar Case Judgment	However, considering compensation and victim reconciliation, the sentence may be moderately reduced. In practice, the combined sentence might range from 1 year and 6 months to 2 years. Based on similar precedent cases, the sentence could be set at 1 year and 6 months of fixed-term imprisonment.

Table 13: Examples of flawed reasoning in the AR task.

Error Types	Example
Confusion between Articles	**Ground Truth**: Article 17 of the Law on the Protection of Women's Rights and Interests: The relevant departments should listen to and adopt criticisms or feasible suggestions regarding the safeguarding of women's rights and interests.**Prediction**:Article 15 of the Law on the Protection of Women's Rights and Interests: The state actively cultivates and selects female cadres, with emphasis on cultivating and selecting female cadres from ethnic minorities.
Content Fabrication	**Ground Truth**: Article 51 of the Veterans Security Law: Veterans who meet the conditions for preferential housing arrangement shall be provided with a combination of market purchase and centralized construction by the military and local governments, which shall be planned and implemented scientifically by the people's government of the resettlement area. **Prediction**: If the unit that receives and resettles demobilized military personnel violates the provisions of these regulations and falls under one of the following circumstances, the competent department of veterans' affairs of the local people's government shall order it to make corrections within a time limit.
Outdated References	**Ground Truth(Revised on February 24, 2017)**: Article 20 of the Red Cross Society Law: The Red Cross shall, in accordance with the law, accept donations of funds and materials from natural persons, legal persons, and other organizations, and issue public welfare donation receipts that are uniformly supervised (printed) by the financial department to the donors. **Prediction(Amended on August 27, 2009)**: The sources and uses of the Red Cross funds shall be reported to the Red Cross Council every year and be subject to inspection and supervision by the People's Government.

in generation tasks (Zheng et al., 2024). Second, we designed structured prompt templates with explicit formatting and structural constraints, which helped reduce hallucinations and improve both consistency and legal formality, consistent with recent empirical findings (de Kinderen and Winter, 2024). Third, we conducted rigorous human review by randomly sampling each batch of generated data and verifying the reasoning process, with particular attention to the correctness of cited legal articles and the professionalism of the language. In summary, while we acknowledge the inherent risk of bias, these measures have effectively mitigated its impact, keeping it within acceptable limits for the purposes of this study.

Table 14: Examples of flawed reasoning in the CPC task.

Type	Example
Citation Factual Inaccuracies	Assuming that Lovern v. Edwards does discuss the reasonable jury standard in the context of summary judgment, the answer would be Yes. If not, then No. Since the user is asking for a Yes/No answer, I need to be certain. Without access to the full case details, I might rely on the common knowledge that Lovern v. Edwards is cited for summary judgment standards, which include the reasonable jury test. Therefore, the answer is likely Yes.
Misunderstanding the Citation	The citation to Gunn v. Minton does not directly support the text. While Gunn addresses federal question jurisdiction, it focuses on whether a state-law claim "arises under" federal law when it implicates a substantial federal issue—not whether federal law creating the cause of action establishes jurisdiction. The text reflects the general rule from cases like Merrell Dow Pharmaceuticals Inc. v. Thompson, not Gunn.

Table 15: Examples of flawed reasoning in the NYSJE task.

Type	Example
Ethical Guidelines Factual Hallucinations	First, I need to recall the judicial ethics rules. Judges are supposed to follow the law, right? Even if they disagree with a law, they have to uphold it unless it's been declared unconstitutional by a higher court. So if a judge believes a statute is unconstitutional, can they just ignore it? Probably not.
Adding Irrelevant Details	But wait, sometimes there's a rule that judges shouldn't appear before other judges in the same county if it could create an appearance of favoritism. However, if both are part-time and their courts are separate, maybe it's okay. The key is whether the Committee has a specific opinion on this.

D.2 Impact of Retrieval Quality

In Table 3, we observe that high-quality supplementary knowledge and contextual information play a crucial role in improving model performance on knowledge-intensive tasks. To further investigate how retrieval quality influences performance, we conducted a controlled experiment where the only variable was the quality of the retrieved documents.

Specifically, we evaluated model performance under three retrieval settings of different quality levels:

High-quality: Gold-reference passages that directly contain the correct answer (ideal retrieval).

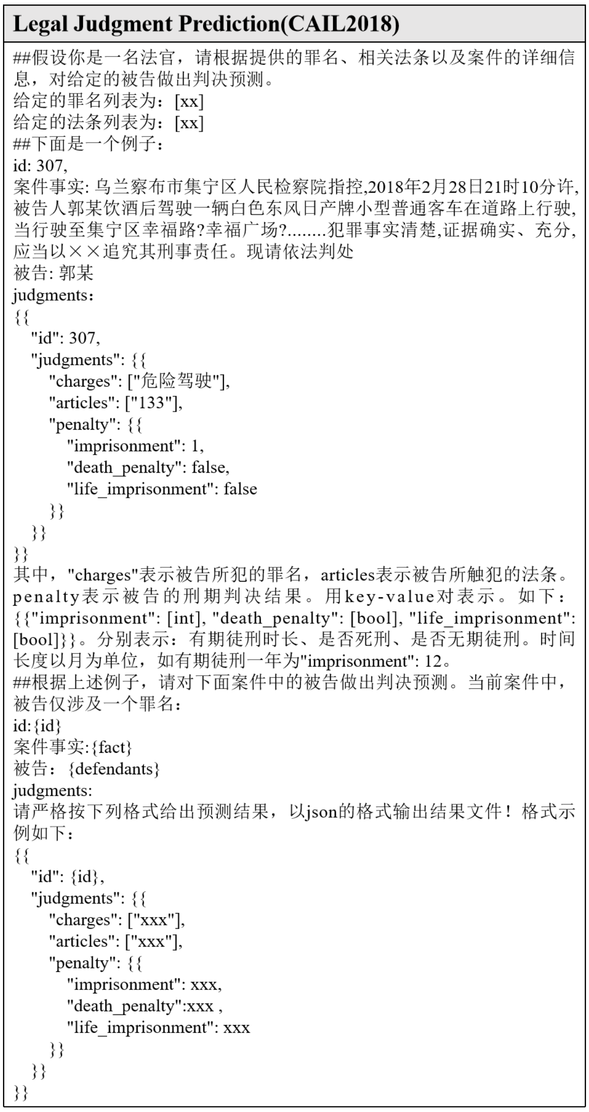
Medium-quality: High-quality content mixed with three unrelated legal articles (moderate noise).

Low-quality: Five randomly selected legal articles unrelated to the input query.

Table 16: Impact of retrieval quality on model performance.

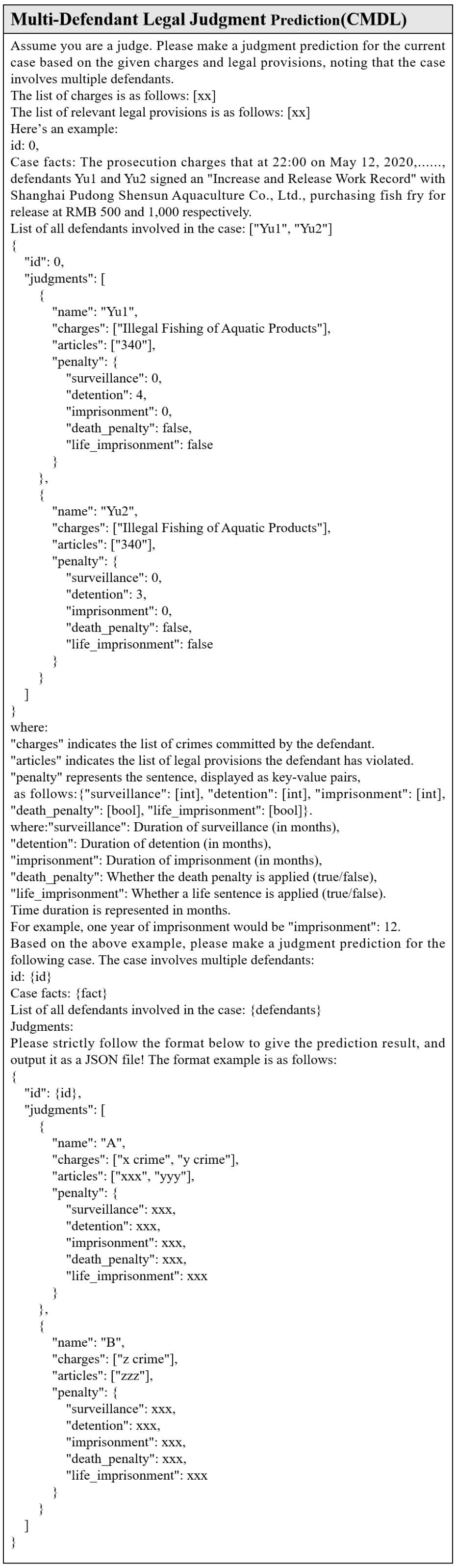
Model	CAIL2018	CMDL	MultiLJP
DeepSeek-R1-Distill-Qwen-14B	72.03%	50.00%	56.50%
with RAG (High-quality)	74.41%	56.94%	66.28%
with RAG (Medium-quality)	70.90%	49.56%	57.23%
with RAG (Low-quality)	67.40%	45.33%	54.19%
DeepSeek-R1	78.00%	68.48%	67.15%
with RAG (High-quality)	80.10%	72.21%	71.49%
with RAG (Medium-quality)	78.37%	68.23%	67.34%
with RAG (Low-quality)	76.17%	65.95%	64.89%

As shown in Table 16, retrieval quality has a substantial impact regardless of whether the base model(DeepSeek-R1-Distill-Qwen-14B) or DeepSeek-R1 is used. When low-quality or noisy content is retrieved, performance drops significantly—even compared with the setting where no external context is provided.



```
Legal Judgment Prediction(CAIL2018)
##假设你是一名法官，请根据提供的罪名、相关法条以及案件的详细信息，对给定的被告做出判决预测。
给定的罪名列表为：[xx]
给定的法条列表为：[xx]
##下面是一个例子：
id: 307,
案件事实: 乌兰察布市集宁区人民检察院指控,2018年2月28日21时10分许,被告人郭某饮酒后驾驶一辆白色东风日产牌小型普通客车在道路上行驶,当行驶至集宁区幸福路?幸福广场?........犯罪事实清楚,证据确实、充分,应当以××追究其刑事责任。现请依法判处
被告: 郭某
judgments：
{{
   "id": 307,
   "judgments": {{
      "charges": ["危险驾驶"],
      "articles": ["133"],
      "penalty": {{
         "imprisonment": 1,
         "death_penalty": false,
         "life_imprisonment": false
      }}
   }}
}}
其中，"charges"表示被告所犯的罪名，articles表示被告所触犯的法条。penalty表示被告的刑期判决结果。用key-value对表示。如下：{{"imprisonment": [int], "death_penalty": [bool], "life_imprisonment": [bool]}}。分别表示：有期徒刑时长、是否死刑、是否无期徒刑。时间长度以月为单位，如有期徒刑一年为"imprisonment": 12。
##根据上述例子，请对下面案件中的被告做出判决预测。当前案件中，被告仅涉及一个罪名：
id:{id}
案件事实:{fact}
被告：{defendants}
judgments:
请严格按下列格式给出预测结果，以json的格式输出结果文件！格式示例如下：
{{
   "id": {id},
   "judgments": {{
      "charges": ["xxx"],
      "articles": ["xxx"],
      "penalty": {{
         "imprisonment": xxx,
         "death_penalty":xxx ,
         "life_imprisonment": xxx
      }}
   }}
}}
```

Figure 7: The Chinese prompt for CAIL2018 dataset.



```
Multi-Defendant Legal Judgment Prediction(CMDL)
Assume you are a judge. Please make a judgment prediction for the current case based on the given charges and legal provisions, noting that the case involves multiple defendants.
The list of charges is as follows: [xx]
The list of relevant legal provisions is as follows: [xx]
Here's an example:
id: 0,
Case facts: The prosecution charges that at 22:00 on May 12, 2020,......, defendants Yu1 and Yu2 signed an "Increase and Release Work Record" with Shanghai Pudong Shensun Aquaculture Co., Ltd., purchasing fish fry for release at RMB 500 and 1,000 respectively.
List of all defendants involved in the case: ["Yu1", "Yu2"]
{
   "id": 0,
   "judgments": [
      {
         "name": "Yu1",
         "charges": ["Illegal Fishing of Aquatic Products"],
         "articles": ["340"],
         "penalty": {
            "surveillance": 0,
            "detention": 4,
            "imprisonment": 0,
            "death_penalty": false,
            "life_imprisonment": false
         }
      },
      {
         "name": "Yu2",
         "charges": ["Illegal Fishing of Aquatic Products"],
         "articles": ["340"],
         "penalty": {
            "surveillance": 0,
            "detention": 3,
            "imprisonment": 0,
            "death_penalty": false,
            "life_imprisonment": false
         }
      }
   ]
}
where:
"charges" indicates the list of crimes committed by the defendant.
"articles" indicates the list of legal provisions the defendant has violated.
"penalty" represents the sentence, displayed as key-value pairs,
 as follows:{"surveillance": [int], "detention": [int], "imprisonment": [int], "death_penalty": [bool], "life_imprisonment": [bool]}.
where:"surveillance": Duration of surveillance (in months),
"detention": Duration of detention (in months),
"imprisonment": Duration of imprisonment (in months),
"death_penalty": Whether the death penalty is applied (true/false),
"life_imprisonment": Whether a life sentence is applied (true/false).
Time duration is represented in months.
For example, one year of imprisonment would be "imprisonment": 12.
Based on the above example, please make a judgment prediction for the following case. The case involves multiple defendants:
id: {id}
Case facts: {fact}
List of all defendants involved in the case: {defendants}
Judgments:
Please strictly follow the format below to give the prediction result, and output it as a JSON file! The format example is as follows:
{
   "id": {id},
   "judgments": [
      {
         "name": "A",
         "charges": ["x crime", "y crime"],
         "articles": ["xxx", "yyy"],
         "penalty": {
            "surveillance": xxx,
            "detention": xxx,
            "imprisonment": xxx,
            "death_penalty": xxx,
            "life_imprisonment": xxx
         }
      },
      {
         "name": "B",
         "charges": ["z crime"],
         "articles": ["zzz"],
         "penalty": {
            "surveillance": xxx,
            "detention": xxx,
            "imprisonment": xxx,
            "death_penalty": xxx,
            "life_imprisonment": xxx
         }
      }
   ]
}
```

Figure 8: The prompt for CMDL dataset.

Multi-Defendant Legal Judgment Prediction(CMDL)

```
##假设你是一名法官，请从给定的罪名和法条中对当前的案件做出判决预测，注意案件中存在多个被告。
 给定的罪名列表为：[xx]
给定的法条列表为：[xx]
##下面是一个例子：
id: 0,
案件事实: 公诉机关指控：2020年5月12日22时许，被告人喻1、喻2在本市内陆水域禁渔期内，携带电捕鱼工具至本市浦东新区万祥镇万五9组路与机耕路交叉口东侧泐马河水域,......,分别以人民币500元、1,000元购买鱼苗用于增殖放流。
该案件涉及的所有被告人姓名列表: ["喻1", "喻2"]
judgments：
{{
   "id": 0,
   "judgments": [
      {{
         "name": "喻1",
         "charges": ["非法捕捞水产品罪"],
         "articles": ["340"],
         "penalty": {{
            "surveillance": 0,
            "detention": 4,
            "imprisonment": 0,
            "death_penalty": false,
            "life_imprisonment": false
         }}
      }},
      {{
         "name": "喻2",
         "charges": ["非法捕捞水产品罪"],
         "articles": ["340"],
         "penalty": {{
            "surveillance": 0,
            "detention": 3,
            "imprisonment": 0,
            "death_penalty": false,
            "life_imprisonment": false
         }}
      }}
   ]
}}
其中，"charges"表示被告人所犯的所有罪名列表，articles表示被告人所触犯的所有法条列表。penalty表示被告人的刑期判决结果。用key-value对表示。如下：
{{"surveillance": [int], "detention": [int], "imprisonment": [int], "death_penalty": [bool], "life_imprisonment": [bool]}},分别表示：拘役时长、管制时长、有期徒刑时长、是否死刑、是否无期徒刑。时间长度以月为单位，如有期徒刑一年为"imprisonment": 12。
##根据上述例子，请你对下面的案件给出判决预测结果：
id:{id}
案件事实:{fact}
该案件涉及的所有被告人姓名列表：{defendants}
judgments:
请严格按下列格式给出预测结果，以json的格式输出结果！格式示例如下：
注意，不要对刑期进行注释。
{{
   "id": {id},
   "judgments": [{{
      "name": "A",
      "charges": ["x罪", "y罪"],
      "articles": ["xxx", "yyy"],
      "penalty": {{
         "surveillance": xxx,
         "detention": xxx,
         "imprisonment": xxx,
         "death_penalty":xxx ,
         "life_imprisonment": xxx
      }}
   }},
   {{
      "name": "B",
      "charges": ["z罪"],
      "articles": ["zzz"],
      "penalty":
       "penalty": {{
         "surveillance": xxx,
         "detention": xxx,
         "imprisonment": xxx,
         "death_penalty":xxx ,
         "life_imprisonment": xxx
      }}
   }}]
}}
```

Figure 9: The Chinese prompt for CMDL dataset.

Multi-Defendant Legal Judgment Prediction(MultiLJP)

```
Assume you are a judge. Based on the given charges and legal provisions, please provide a judgment for multiple defendants in the case, including the charges, legal provisions, and sentence.
The list of charges and legal provisions is as follows: XXX
id: {id}
Case facts: {fact}
Defendant list: {defendants}
Please strictly follow the format below to give the prediction result, and ensure the case id is included in the prediction result! The format example is as follows:
{
   "id": {id},
   "judgments": {
      "[Defendant A]": {
         "accusations": ["xx"],
         "laws": [xx],
         "term": "detention for xx months | imprisonment for xx months/year"
      },
      "[Defendant B]": {
         "accusations": ["xx"],
         "laws": [xx],
         "term": "imprisonment for xx months/year | detention for xx months"
      }
   }
}
Chinese:
###
假设你是一名法官，请你从给定的罪名和法条中，对案件中的多个被告给出判决，包括罪名，法条和刑期。
给定的罪名和法条如下： XXX
###
id:{id}
案件事实:{fact}
被告列表:{defendants}
##
请严格按下列格式给出预测结果，以json的格式输出结果，并确保案件的id存在于预测结果中！格式示例如下：
{{
   "id": {id},
   "judgments": {{
      "[被告A]": {{
         "accusations": ["xx"],
         "laws": [xx],
         "term": "拘役xx个月|有期徒刑xx个月/年"
      }},
      "[被告B]": {{
         "accusations": ["xx"],
         "laws": [xx],
         "term": "有期徒刑xx个月/年|拘役xx个月"
      }}
   }}
}}
```

Figure 10: The prompt for MultiLJP dataset.

Multi-Defendant Charge Prediction

```
Task Description:
Assume you are a judge. Based on the given charges, make a judgment on the charges committed by the defendants in the case. If a defendant is involved in multiple charges, choose the more serious or applicable charge. Note that the case involves multiple defendants.
The list of charges is as follows: [xx]
Input:
id: {id}
Case facts: {facts}
Defendant list: {defendants}
Output:
Please strictly follow the JSON format below to output the prediction result and ensure the id is included in the result.
{
  "id": {id},
  "judgments": [
    {
      "subject": "Defendant 1",
      "charge": "Charge committed by Defendant 1"
    },
    {
      "subject": "Defendant 2",
      "charge": "Charge committed by Defendant 2"
    }
  ]
}
Chinese:
###
任务说明：
假设你是一名法官，请你从给定的罪名中，对案件中的被告所犯的罪名做出判决，如果被告涉及多个罪名，则选择更严重或更符合的一个罪名即可。注意案件中存在多个被告。
给定的罪名如下：[xx]
###
输入：
id:{id}
案件事实:{facts}
被告列表:{defendants}
###
输出：
请严格按下列JSON的格式输出预测结果，并确保id存在于预测结果里面。
{{
  "id": {id},
  "judgments": [
    {{
      "subject": "被告1",
      "charge": "被告1所犯的罪名"
    }},
    {{
      "subject": "被告2",
      "charge": "被告2所犯的罪名"
    }}
  ]
}}
```

Figure 11: The prompt for MUD dataset.

Multi-segment Legal Reading Comprehension

```
Please answer the elements in the question based on the current case information and present them in a list format. Each element in the answer should be in the form of a "string".
Input:
caseid: {caseid}
Fact description: {context}
Question: {question}
Here are a few examples:
Question: What expenses did the plaintiff incur?
Answer: ['Loan of 210,000 yuan', 'Lawyer's fee of 4,000 yuan']
Question: When was the civil ruling made, and what is the case acceptance fee?
Answer: ["January 9, 2012", "10,800 yuan"]
Please strictly follow the format below to give the answer, and ensure the id is included in the answer. The format is as follows:
{
  "id": "{caseid}",
  "answer": []
}
Chinese:
请基于当前案件信息，回答问题中的要素，并将其以列表的格式呈现。答案中的每个要素都应以"字符串"形式呈现。
caseid：{caseid}
事实描述：{context}
问题：{question}
下面是几个例子：
问题：原告支出了哪些款项。
answer: ['21万元借款','律师费4000元']
问题：法院何时作出的民事裁定书，案件受理费为多少。
answers: ["2012年1月9日", "10800元"]
请严格按下列格式给出答案，以json的格式输出，并确保id存在于答案中，格式如下：
{{
  "id": "{caseid}",
  "answer": []
}}
```

Figure 12: The prompt for MSLRC dataset.

Controversial Focus Extraction

```
Task Description:
Assume you are a judge. Based on the given dispute focal points dictionary, extract the specified number of dispute focal points from the case facts and present them in a list of indices, such as [1, 2, 3].
The dispute focal points dictionary is as follows:
Dispute focal points dictionary: {xx}
Case information is as follows:
Case ID: {id}
Case facts: {fact}
Number of dispute focal points: {num}
Output:
Please strictly follow the format below to provide the result in JSON format, ensuring that the case id is included in the result. The format is as follows:
{{
  "id": "{id}",
  "answer": []
}}
Chinese:
任务说明： 假设你是一名法官，请你根据给定的争议焦点字典，从案件事实中提取指定个数的争议焦点，并以列表的方式给出争议焦点的索引，如[1,2,3]。
给定的争议焦点字典如下： 争议焦点字典：{xx}
案件信息如下：
案件id:{id}
案件事实： {fact}
争议焦点个数： {num}个

输出：
请严格按下列格式给出结果，以json的格式输出结果，并确保案件的id存在于结果中！格式示例如下：
{{
"id": "{id}",
"answer": []
}}
```

Figure 13: The prompt for CFE dataset.

Interactive Argument-Pair Extraction

```
You are a legal expert, and you will perform the task of argument pair extraction. Given a claimant's argument, you need to identify one answer from the five defendant’s candidate arguments that could form an interactive argument pair with it. The interactive argument could either be consistent or partially consistent, but in most cases, one argument will be in dispute with another.
Case information is as follows:
Claimant’s argument: {sc}
Five defendant’s candidate arguments:
1: {bc1}
2: {bc2}
3: {bc3}
4: {bc4}
5: {bc5}
Note: Only provide the number of the argument, which will be one of the following: 1, 2, 3, 4, or 5. Do not output any other content.
Please strictly follow the format below to provide the result in JSON format. The format is as follows:
{
  "id": "{id}",
  "answer": ""
}
Chinese:
你是一名法律专家,您将执行论点对挖掘的任务，即给定一个诉方观点，识别五个辩方候选论点中可以与之形成互动论点对的一个答案。互动的论点可能是一致的或部分一致的，但在大多数情况下，一个论点会与另一个论点发生争议。
#下面是案件信息:
诉方论点: {sc}
五个辩方候选论点:
1:{bc1}
2:{bc2}
3:{bc3}
4:{bc4}
5:{bc5}
注意：仅仅给出一个论点的编号，即：1、2、3、4、5中的一个，不要输出其他内容！
请严格按下列格式给出预测结果，以json的格式输出结果！格式示例如下：
{{
  "id": "{id}",
  "answer": ""
}}
```

Figure 14: The prompt for IAPE dataset.

Article Recitation

```
Based on China's legal and regulatory framework, please answer the following question by providing only the text of the relevant legal provisions.
{id}
{input}
Please strictly output the prediction result in the following JSON format:
{
  "id": "{id}",
  "answer": ""
}
Chinese:
请依据中国的法律法规体系，回答以下问题，只需直接给出法条内容。

{id}
{input}

请严格按下列格式给出预测结果，以json的格式输出结果文件！格式如下：
{{
  "id": "{id}",
  "answer": ""
}}
```

Figure 15: The prompt for AR dataset.

Judicial Examination

```
You are a judge and need to answer the question based on the provided legal background and options. The answer must be a list of one or more options. Please answer the current question as follows:
id:{id}
statement: {statement}
option_list: {option_list}
Please strictly follow the following format to give the prediction result and output it in json format! Example format:
{
  "id": "{id}",
  "answer": []
}
Chinese:
你是一名法官，需要根据提供的法律背景和选项回答问题。答案必须是一个或多个选项的列表。请你对当前问题做出回答：
id:{id}
statement: {statement}
option_list: {option_list}
请严格按下列格式给出预测结果，以json的格式输出结果！格式示例如下：
{{
  "id": "{id}",
  "answer": []
}}
```

Figure 16: The prompt for JE dataset.

Legal_Reasoning_Causality

```
Do the following opinion excerpts rely on statistical evidence? Answer Yes or No.
Excerpt: {text}
Answer:
Please output the Answer in the following JSON format and ensure that the 'id' is included in the response.
{{
  "id": "{id}",
  "answer": ""
}}
Chinese:
以下意见摘录是否依赖统计证据？回答是或否。
摘录：{文本}
答案:
请按照以下JSON格式输出答案，并确保在响应中包含'id'。
{{
"id": "{id}",
"answer": ""
}}
```

Figure 17: The prompt for LRC dataset.

Citation_Prediction_Classification

```
Can the case can be used as a citation for the provided text?  Answer Yes or No.
Text: {text}
Citation: {citation}
Supportive?
Answer:
Please output the Answer in the following JSON format and ensure that the 'id' is included in the response.
{{
  "id": "{id}",
  "answer": ""
}}
Chinese:
案例可以作为所提供文本的引用吗？ 回答是或否。
文本：{文本}
引用：{引用}
支持吗？
答案:
请按照以下 JSON 格式输出答案，并确保在响应中包含“id”。
{
"id": "{id}",
"answer": ""
}
```

Figure 18: The prompt for CPC dataset.

NYS_Judicial_Ethics

```
Imagine your are the New York State Unified Court System Advisory Committee on Judicial Ethics. You've received the following question(s). Answer them as either "Yes" or "No".
Question: {text}
Answer:
Please output the Answer in the following JSON format and ensure that the 'id' is included in the response.
{{
   "id": "{id}",
   "answer": ""
}}
Chinese:
想象你是纽约州统一法院系统司法道德咨询委员会。你收到了以下问题。回答“是”或“否”。
问题：{文本}
答案:
请按照以下 JSON 格式输出答案，并确保在响应中包含“id”。
{
"id": "{id}",
"answer": ""
}
```

Figure 19: The prompt for NYSJE dataset.

Sara_Numeric

```
Answer the following questions.
Please output the answer in the following JSON format and ensure that the 'id' is included in the response.
{{
   "id": "{id}",
   "answer": ""
}}
Statute: {statute}
Description: {description}
Question: {question}. State the amount first.
Answer:
Chinese:
回答以下问题。
请按照以下 JSON 格式输出答案，并确保在响应中包含“id”。
{
"id": "{id}",
"answer": ""
}
法规：{法规}
描述：{描述}
问题：{问题}。首先说明金额。
答案:
```

Figure 20: The prompt for SARA_N dataset.

Sara_Entailment

```
Determine whether the following statements are entailed under the statute.
Statute: {statute}
Description: {description}
Statement: {question}
Answer:
Then output the answer in the following format:
{{
   "id": "{id}",
   "answer": ""Entailment" or "Contradiction""
}}
please check the output format carefully and make sure the output is in the correct format.
Chinese:
确定以下陈述是否在法规下必然成立。
法规：{法规}
描述：{描述}
陈述：{问题}
答案:
然后按照以下格式输出答案:
{
"id": "{id}",
"answer": "必然成立" 或 "矛盾"
}
```

Figure 21: The prompt for SARA_E dataset.

Scalr

```
Given the following question presented in a court case, select the most relevant holding.
Please output the answer in the following JSON format and ensure that the 'id' is included in the response.
{{
   "id": "{id}",
   "answer": ""
}}
Question: {question}
Choices:
0: {choice_0}
1: {choice_1}
2: {choice_2}
3: {choice_3}
4: {choice_4}
Answer:
Chinese:
鉴于以下在法庭案件中提出的问题，请选择最相关的裁决。
请按照以下 JSON 格式输出答案，并确保在响应中包含 'id'。
{
"id": "{id}",
"answer": ""
}
问题: {question}
选项:
0: {choice_0}
1: {choice_1}
2: {choice_2}
3: {choice_3}
4: {choice_4}
答案:
```

Figure 22: The prompt for Scalr dataset.

LAR

```
You will be provided with the introductory Facts in a European Court of Human Rights (ECHR) case, an excerpt of arguments from that case and several possible continuations of these arguments. Your task is to determine which continuation accurately extends the original argument.
There is the following information:
Facts: {facts}
Preceding arguments: {preceding_arguments}
Continuation options:
A: {choice_1}
B: {choice_2}
C: {choice_3}
D: {choice_4}
Please output the Answer in the following JSON format and ensure that the 'id' is included in the response.
{{
   "id": "{id}",
   "answer": ""
}}
Chinese:
您将获得欧洲人权法院（ECHR）案件的引入事实、该案件的一段辩论以及几个可能的辩论延续。您的任务是确定哪个延续准确地延伸了原始辩论。
以下是有关信息:
事实：{事实}
前面的论点：{前面的论点}
延续选项:
A：{选择_1}
B：{选择_2}
C：{选择_3}
D：{选择_4}
请按照以下 JSON 格式输出答案，并确保在响应中包含'id'。
{
"id": "{id}",
 "answer": ""
}
```

Figure 23: The prompt for LAR dataset.