PLAWBENCH: A Rubric-Based Benchmark for Evaluating LLMs in Real-World Legal Practice

Anonymous ACL submission

Abstract

As large language models (LLMs) are increasingly applied to legal domain-specific tasks, evaluating their ability to perform legal work in real-world settings has become essential. However, existing legal benchmarks rely on simplified and highly standardized tasks, failing to capture the ambiguity, complexity, and reasoning demands of real legal practice. Moreover, prior evaluations often adopt coarse, single-dimensional metrics and do not explicitly assess fine-grained legal reasoning. To address these limitations, we introduce **PLAWBENCH**, a **P**ractical **Law Bench**mark designed to evaluate LLMs in realistic legal practice scenarios. Grounded in real-world legal workflows, PLAWBENCH models the core processes of legal practitioners through three task categories: public legal consultation, practical case analysis, and legal document generation. These tasks assess a model's ability to identify legal issues and key facts, perform structured legal reasoning, and generate legally coherent documents. PLAWBENCH comprises 850 questions across 13 practical legal scenarios, with each question accompanied by expert-designed evaluation rubrics, resulting in approximately 12,500 rubric items for fine-grained assessment. Using an LLM-based evaluator aligned with human expert judgments, we evaluate 10 state-of-the-art LLMs. Experimental results show that none achieves strong performance on PLAWBENCH, revealing substantial limitations in the fine-grained legal reasoning capabilities of current LLMs and highlighting important directions for future evaluation and development of legal LLMs. Data is available at: https://anonymous.4open.science/r/PLawbench-B524/.

1 Introduction

With the rapid advancement of AI technologies, large language models (LLMs) have been increasingly applied to a wide range of domain-specific applications, such as legal (Chen et al., 2024b; Padiu et al., 2024), medical (Zhou et al., 2023), and finance (Li et al., 2023). In these knowledge-intensive fields, real-world problems often require the integration of extensive prior knowledge together with complex logical reasoning. Enabling LLMs to handle professional tasks in such domains, therefore, necessitates the coordinated use of multiple capabilities, including knowledge retrieval (Pipitone and Alami, 2024), reasoning (Yao et al., 2025), planning (Li et al., 2025a), and decision making (Chen et al., 2024a).

In the legal domain, LLMs have been explored as tools to assist with real-world legal tasks traditionally performed by legal practitioners. Practical legal scenarios are highly diverse and complex, requiring the ability to retrieve and apply legal knowledge, identify and clarify legally relevant facts, plan multi-step problem-solving strategies, conduct structured legal reasoning (Shi et al., 2025), and verify and analyze conclusions (American Bar Association, 1992; Doyle and Tucker, 2025). This raises a fundamental question: **To what extent can LLMs handle complex legal tasks and reason like human practitioners?**

Prior research has attempted to evaluate LLMs on legal tasks (Cui et al., 2023; Hu et al., 2025; Liu et al., 2025b; Fei et al., 2024; Li et al., 2025b). However, the construction of these benchmarks often fails to reflect the practical demands of realistic legal workflows. Specifically, existing work suffers from the following three limitations:

① **Lack of Realistic Task Construction.** Some previous benchmarks are directly derived from judicial or bar examination questions (e.g., LAIW (Dai et al., 2025), LEXAM (Fan et al., 2025)), while others contain tasks that are deliberately simplified (e.g., LegalAgentBench (Li et al., 2025a), Legalbench (Guha et al., 2023)). In real-world legal settings, user queries are frequently ambiguous, in-

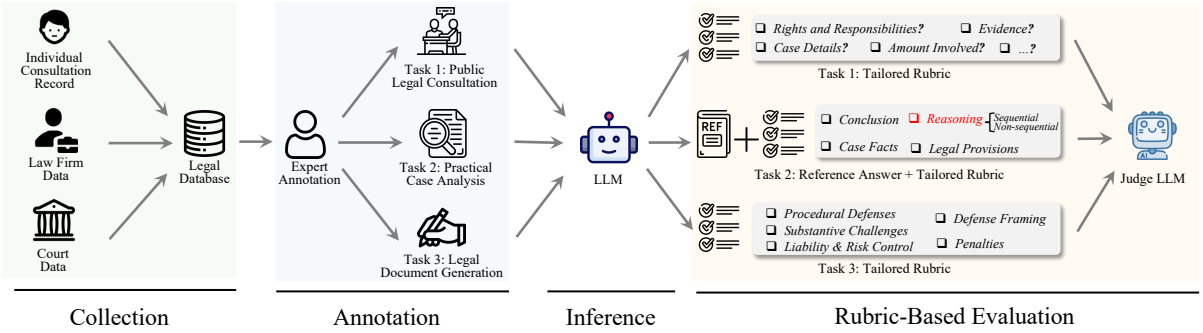

Figure 1: Overall framework of PLAWBENCH illustrates a four-step pipeline: collecting multi-source legal data, expert annotation into three task types, LLM-based inference on these tasks, and rubric-based evaluation of the LLM outputs by a judge model.

complete, or strategically framed, with key facts omitted or obscured. Only by incorporating such characteristics can evaluations more accurately capture LLM performance when acting as legal practitioners.

② **Insufficient Modeling of Legal Reasoning.** Logical and structured legal reasoning steps are essential when large language models handle legal problems. Prior benchmarks claim to assess reasoning abilities. LexEval (Li et al., 2024a) evaluates "Logic Inference" capability while their designs overlook the distinctive nature of legal reasoning. Simple syllogistic reasoning (Deng et al., 2023a) or IRAC (Issue, Rule, Application, and Conclusion)-style (Jang et al., 2025) evaluations primarily capture surface structure rather than substantive legal inference (Zhang et al., 2025a). Even recent benchmarks, such as PRBENCH (Akyürek et al., 2025), rely on general-purpose scoring frameworks that are not tailored to the unique reasoning mechanisms of the legal domain. As a result, these approaches fail to adequately assess the complex, context-dependent reasoning required in real legal practice.

③ **Coarse and Uniform Evaluation Metrics.** Moreover, existing evaluations typically rely on coarse criteria such as whether the final conclusion is correct, whether relevant statutes are cited, or whether a syllogistic structure is followed (Fei et al., 2024; Li et al., 2024a; Wald, 2018). However, such uniform evaluation metrics fail to capture the diversity and complexity of legal practice across different scenarios. Fine-grained and case-specific evaluation rubrics are essential for realistically assessing legal practitioner LLMs.

Due to these limitations, existing benchmarks have not achieved a comprehensive, in-depth, or a realistic assessment of LLM capabilities in legal practice. To address this gap, we propose a **P**ractical **Law Bench**mark (**PLAWBENCH**) designed to evaluate LLMs acting as legal practitioners under realistic, knowledge-intensive, and reasoning-driven conditions. This work has three main contributions:

❶ **Authentic Simulation of Legal Practice.** We evaluate LLMs using a hierarchical framework adapted from authentic legal workflows, comprising three levels: *Public Legal Consultation*, *Practical Case Analysis*, and *Legal Document Generation* (see Figure 1) [1]. To mirror the cognitive challenges of real-world practice, we adapt tasks from actual cases and deliberately incorporate realistic noise — such as vague queries, emotional narratives, and omitted facts. This design moves beyond idealized benchmarks to rigorously test whether models can distill relevant legal facts from the ambiguity inherent in professional environments.

❷ **Fine-Grained Reasoning Steps.** Beyond evaluating the final conclusions, our benchmark explicitly incorporates fine-grained legal reasoning steps into task design and evaluation. This allows us to examine whether LLMs can perform multi-stage legal reasoning, including issue identification, fact clarification, legal analysis, and conclusion validation, rather than relying on shallow pattern matching or surface-level reasoning.

❸ **Task-Specific Rubrics.** Our framework employs personalized, task-specific rubrics to assess

[1] Adapted from Maughan and Webb (Maughan and Webb, 2005): *Consultation* corresponds to "Interviewing" (Ch.5), *Case Analysis* to "Case Management" (Ch.10), and *Document Generation* to "Drafting" (Ch.7-8).

substantive legal reasoning beyond mere outcomes. Legal experts follow a two-stage annotation process: first defining a reasoning-oriented framework for each task type, then tailoring specific criteria to individual case scenarios. This approach ensures evaluation is both principled and context-sensitive, enabling a fine-grained and realistic assessment of LLM performance in practical settings.

2 Related Work

Existing research has proposed a variety of benchmarks to evaluate AI systems in legal domains, primarily focusing on models' ability to understand and apply legal knowledge. Early efforts assess legal AI systems through tasks such as legal information extraction and legal reasoning. Representative benchmarks include JEC-QA (Zhong et al., 2020), which evaluates multi-hop reasoning based on Chinese bar examination questions; CaseHOLD (Zheng et al., 2024), which focuses on identifying legal holdings in U.S. appellate decisions; (Zhang et al., 2025b), which evaluates the identification of overruling relationships in U.S. Supreme Court cases; and SARA (Blair-Stanek et al., 2023), which examines statutory reasoning in U.S. tax law. Many of these benchmarks adopt syllogistic reasoning structures or legal writing frameworks such as IRAC.

More recently, comprehensive benchmarks have been introduced to provide multi-dimensional evaluations of legal AI systems. For example, LawBench (Fei et al., 2024) evaluates legal LLMs across cognitive levels, including memorization, understanding, and application; LegalBench (Guha et al., 2023) comprises a wide range of tasks spanning multiple forms of legal reasoning; LexEval (Li et al., 2024a), the largest Chinese legal benchmark to date, covers diverse legal tasks at scale; and LAiW (Dai et al., 2025) explicitly structures legal reasoning into layered components aligned with legal practice logic.

Despite their contributions, existing benchmarks often fail to reflect the complexities of real-world legal practice. First, tasks like statutory memorization or exam-based QA (e.g., LawBench (Fei et al., 2024)) diverge significantly from authentic scenarios where practitioners must distill relevant facts from ambiguous, incomplete client narratives. Second, current benchmarks typically isolate interdependent tasks — such as separating statutory retrieval from judgment prediction — ignoring the holistic reasoning chain required in practice. Furthermore, evaluation metrics for document generation (e.g., ROUGE-L) fail to capture professional legal standards. In contrast, our benchmark integrates three representative authentic tasks with fine-grained, expert-designed rubrics, ensuring comprehensive and practically relevant evaluation. A comprehensive review of related work is provided in Appendix A.

Dimension	What It Evaluates
Issue & Fact Identification	Identification of legal issues and material facts
Legal Reasoning	Logical inference connecting facts and legal rules
Legal Knowledge Application	Context-aware application of legal provisions
Procedural & Strategic Awareness	Procedural compliance and litigation strategy
Claim & Outcome Construction	Formulation of claims, defenses, and remedies
Professional Norms & Compliance	Adherence to legal and ethical standards

Table 1: Rubric dimensions defined based on practical legal tasks.

3 PLawBench

3.1 Framework

Overall Framework As illustrated in Figure 1, we organize legal tasks into three hierarchical levels based on the real-world workflow of legal practitioners: public legal consultation, practical case analysis, and legal document generation. These three tasks are designed to evaluate distinct but complementary capabilities of LLM-based agents: identifying legal issues and clarifying key facts, analyzing disputes through legal reasoning, and summarizing and presenting legal claims in a structured and formalized manner, respectively.

Rubric Construction Our rubric design adopts a risk-tiered perspective. It is grounded in principles emphasized by the EU AI Act's (Act, 2024) risk classification framework, AI auditing standards, and legal risk management practices: the risks posed by AI systems must be evaluated in relation to specific tasks, concrete contexts, and affected stakeholders, rather than relying solely on abstract accuracy metrics (Hagan, 2024; Calloway, 2025).

In legal settings, traditional notions of "accuracy" are insufficient to capture task complexity. In many scenarios, information may be factually correct yet operationally unusable, legally accurate but high-risk in terms of rights implications, or correct but lacking essential procedural guidance — all of which can result in substantial harm to users.

Accordingly, we abstract task-level rubric items into six unified evaluation dimensions, each corresponding to a fundamental component of legal practice: (1) Issue and Fact Identification, (2) Legal

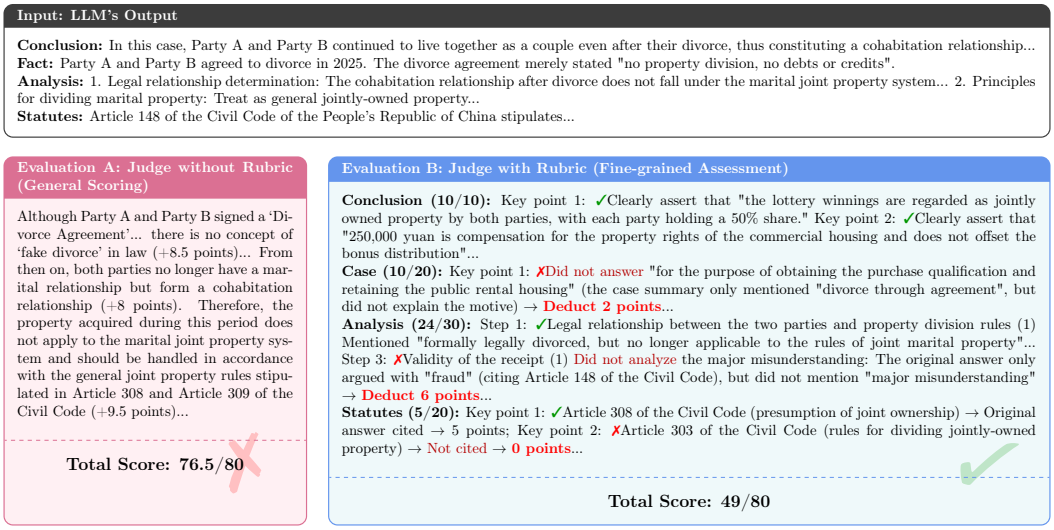

Figure 2: A contrasting example: rubric-based approach (**Evaluation B**) can identify situations that appear accurate on the surface but are actually flawed in the reasoning process, while rubric-free approach (**Evaluation A**) cannot do this, potentially exposing users to significant legal risks in practice.

Reasoning, (3) Legal Knowledge Application, (4) Procedural and Strategic Awareness, (5) Claim and Outcome Construction, and (6) Professional Norms and Compliance as shown in Table 1. These dimensions are designed to capture both substantive legal reasoning and practical decision-making, rather than surface-level task performance. Building on the six unified evaluation dimensions, we map them onto specific legal task settings and integrate them into rubric design to ensure that risk assessment aligns with task-specific legal risk points.

As shown in Figure 2, this unified rubric design allows PLAWBENCH to move beyond fragmented task evaluations and provides a principled framework for assessing whether LLMs exhibit the core competencies required in real-world legal practice.

3.1.1 Public Legal Consultation Task

Task Definition The public legal consultation tasks primarily evaluate a model's ability to identify legal issues and clarify key facts. In real-world consultations, users often describe their situations ambiguously, omit critical information, or intentionally conceal unfavorable details. In some cases, emotional expressions overlap with criminal law terminology, for example, users may claim that they were "defrauded", even when the statutory elements of fraud under criminal law are not satisfied. In other scenarios, users' subjective perceptions diverge substantially from legally accurate classifications; for instance, in cases involving illegal fundraising, users may believe they were "scammed", while the applicable criminal offense is illegal fundraising rather than fraud.

Consequently, effective legal consultation relies on iterative questioning to clarify the user's intent and reconstruct legally relevant facts. This requires models to demonstrate the ability to actively elicit key information by posing appropriate follow-up questions, gradually identifying the facts that materially affect legal outcomes.

Rubric Compared to traditional legal QA tasks, which are based on pre-structured, exam-style questions and primarily emphasize answer correctness, the public legal consultation task is designed to assess a model's ability to identify legal risks that emerge in real-world practice. Such risks must be inferred from the model's performance in issue and fact identification, legal reasoning, and legal knowledge application when responding to user narratives that are ambiguous, emotionally charged, or incomplete. These deliberately omitted facts represent the core legal risk points in real consultation scenarios, and the model must identify and clarify them to detect and manage such risks.

To enable systematic evaluation, we design task-specific rubrics focusing on these critical aspects. Each rubric targets a indispensable fact that is not explicitly stated in the prompt, must be clarified through inquiry, and plays a decisive role in legal qualification or the allocation of rights and obligations. A detailed description of the rubric design

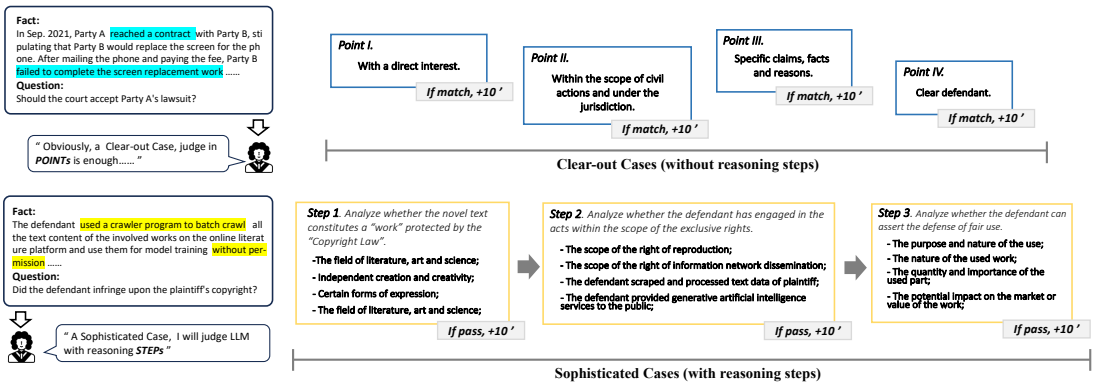

Figure 3: Two legal reasoning modes: clear-out cases judged directly by matching key points (top), and sophisticated cases evaluated through step-by-step legal reasoning over intermediate questions (bottom).

and examples is provided in the Appendix C.1.

3.1.2 Practical Case Analysis Task

Task Definition The practical case analysis tasks focus on evaluating LLMs' capabilities of analyzing disputes through legal reasoning. We assess large language model from two perspectives.

First, all questions in this task are simulated with scenarios that users are likely to encounter in practice. To ensure comprehensive coverage, we construct a legal knowledge framework spanning multiple legal domains. Second, given the importance of statutes of limitations and legal theory application in practice, we introduce two dedicated scenarios to explicitly evaluate models' capabilities in handling temporal legal constraints and doctrinal reasoning.

Recognizing the diversity and complexity of real-world legal consultation demands, we adopt a practice-oriented classification system. We systematically analyze multiple layers of reference materials, including official business classification systems of top-tier law firms in China, categorization standards from global legal service rankings, and business structures used by consumer-facing legal consultation platforms (e.g., China Legal Service Network[2]).

Rubric Unlike law-retrieval tasks that merely restate statutory provisions or legal judgment prediction tasks that output opaque conclusions, legal case analysis requires coherent integration of Legal Reasoning, Legal Knowledge Application, and Claim Outcome Construction. In real legal practice, the factual background of a case is often relatively clear, yet the application of legal rules, element assessment, allocation of rights and obligations, and procedural requirements remains highly complex.

In designing the rubrics, we require models to grasp the conditions under which legal rules apply, identify the facts that are legally determinative, and preserve both logical and doctrinal coherence throughout the reasoning process.

To move beyond overly formalistic legal reasoning evaluation, we revise our rubric design to emphasize substantive legal reasoning. Recognizing that law is fundamentally a reasoning-driven practice, we adapt classical legal syllogism into a framework consisting of Legal Statutes, Case Facts, Reasoning, and Conclusion. While Legal Statutes and Case Facts function as the major and minor premises, respectively, we highlight the critical role of the reasoning step that bridges abstract legal concepts and concrete factual descriptions.

In practice, legal premises rarely align directly with case facts, requiring explicit deductive justification to subsume specific facts under legal concepts. This deductive reasoning step is therefore treated as a core evaluative component and assigned a higher weight than traditional elements. As shown in Figure 3, scoring strictly enforces logical coherence: credit for reasoning is granted only when the response follows a correct logical sequence. In contrast, conclusions, facts, and legal provisions are evaluated with lower weights.

Finally, to better serve lay users, we adopt an output structure that presents the claim first, followed by the complete reasoning process, ensuring transparency and persuasive clarity. A detailed explanation of the reasoning framework and scoring criteria is provided in the Appendix C.2.

[2] https://www.12348.gov.cn/#/homepage

3.1.3 Legal Document Generation Task

Task Definition The legal document generation tasks evaluate the ability to summarize and present legal claims in a structured legal format. Unlike prior benchmarks(Xiao et al., 2018) that provide structured case descriptions and directly request document drafting, our dataset emphasizes simulation of the full real-world legal workflow.

Specifically, we intentionally increase the complexity of legal elements by introducing multiple related parties, ambiguous responsibility boundaries, layered rights and obligations, and scenarios involving the application of both legacy and newly enacted laws. In addition, we innovate in prompt presentation by replacing standardized case descriptions with disorganized narratives from the client's perspective. These narratives incorporate unreasonable demands, incorrect legal terminology, and redundant or misleading information to approximate real-world lawyer-client interactions.

Under this setting, evaluation extends beyond document drafting. Models are required to complete the entire workflow of information filtering, legal relationship analysis, litigation strategy formulation, and document drafting. This enables a comprehensive assessment of the model's problem-solving ability in realistic legal practice scenarios.

Rubric The legal document generation task primarily corresponds to the dimensions of Claim & Outcome Construction, Procedural & Strategic Awareness, and Normative Compliance in our framework.

In designing the rubrics, the model must identify legally relevant facts and claims, make sound procedural and strategic judgments, construct structured and persuasive legal claims, and ensure that the document adheres to required formal and normative standards.

A detailed specification of rubric components and penalty rules is provided in the Appendix C.3.

3.2 Dataset

3.2.1 Data Procurement and Filtering

We present a fine-grained annotated dataset derived from two primary sources: expert-edited judgment documents from China Judgments Online[3], individual consultation records, and complex cases curated by leading Chinese law firms. These open-ended cases represent more intricate disputes that exceed the complexity of standard exam-based questions. Following initial collection, we performed a rigorous screening and anonymization process to finalize 850 representative cases covering diverse legal scenarios.

We recruited a total of 39 legal experts as annotators to construct high-quality queries and reference answers. These experts include legal practitioners from law firms and Ph.D. candidates from top-tier universities. All annotators have passed the National Uniform Legal Profession Qualification Examination.

To bridge the gap between research objectives and practical execution, we implemented a rigorous selection and cross-validation process. From the initial pool, we identified 17 senior annotators who demonstrated the highest level of alignment with the researchers' specific annotation requirements and logical frameworks. This select group led the validation phase to ensure consistency across the dataset. Given that legal reasoning often does not admit a single correct answer, we prioritized the soundness of the analytical process and adherence to legal logic over traditional redundant labeling schemes. This high-intensity process required an average of three hours per data instance. Detailed information regarding the annotation process is provided in Appendix B.1.

3.2.2 Dataset Statistics

The benchmark comprises 850 questions covering a wide range of legal functions. Table 2 presents a summary of key statistics across the three legal tasks, including the number of questions, average prompt length, and the distribution of rubric items. These statistics illustrate the differences in scale, task complexity, and evaluation granularity among the tasks. A detailed description of the dataset composition and additional statistics is provided in Appendix B.

Metric	Task 1 PLC	Task 2 PCA	Task 3 LDG
Total Question Num.	60	750	40
Avg. Prompt Length	1171	692	2574
Rubrics Num.(total)	550	11000	900
Rubrics Num.(avg.)	21	14.1	23

- All data has been annotated by human experts.

Table 2: Dataset Statistics.

[3] https://wenshu.court.gov.cn/

Model	Overall	Task 1: Public Legal Consultation	Task 2: Practical Case Analysis					Task 3: Legal Document Generation		
			Conclusion	Facts	Reasoning	Statute	Average	Defendant	Plaintiff	Average
Claude-Sonnet-4.5	64.10	67.26	67.53	88.41	62.95	48.44	**67.37**	58.80	54.28	56.54
Gemini-2.5-Pro	62.54	71.28	69.59	78.29	63.94	43.49	63.82	58.25	50.88	54.57
Gemini-3.0-Pro-Preview	65.16	65.27	**71.75**	78.73	**64.64**	45.98	64.99	65.30	**65.41**	**65.36**
GPT-4o-20240806	39.26	47.10	54.83	67.90	34.90	16.15	42.00	33.88	25.05	29.47
GPT-5-0807-Global	65.87	**83.67**	66.91	87.32	61.99	35.75	63.93	66.50	47.98	57.24
GPT-5.2-1211-Global	**68.13**	79.57	69.93	88.26	60.38	**48.59**	66.37	**68.58**	58.25	63.42
DeepSeek-V3.2	58.43	70.01	63.99	86.49	59.23	43.95	63.93	48.74	34.33	41.54
Kimi-K2	60.64	67.27	65.55	81.96	58.56	43.28	62.16	60.71	46.65	53.68
Qwen3-235b-Think	55.74	60.64	64.83	89.23	59.97	43.49	64.84	44.37	30.26	37.32
Qwen3-Max	64.75	75.76	67.52	**90.97**	62.75	45.10	67.17	57.43	49.33	53.38

Table 3: Overall performance(%) of evaluated models across three primary tasks. *Task 2 Average* denotes the aggregate scoring rate derived from the specific rubrics of all four sub-dimensions. The *Overall* score is calculated as a weighted mean of Tasks 1, 2, and 3 (with weights 2:5:3, reflecting the number of questions in each task). Best results in each category are **bolded**.

4 Experiment

4.1 Experimental Setup

We conducted experiments across all three tasks: Public Legal Consultation (PLC), Practical Case Analysis (PCA), and Legal Document Generation (LDG). Following previous work (Li et al., 2025a), we selected 10 representative LLMs, covering the GPT (Achiam et al., 2023) , Gemini (Comanici et al., 2025), Claude (Anthropic, 2025), Qwen (Yang et al., 2025), DeepSeek (Liu et al., 2025a) , and Kimi (Team et al., 2025) series. All LLMs are accessed through API calls. We utilized default hyperparameters for reproducibility, solely adjusting the max_token limit to 16k / 32k to ensure response completeness. To ensure structured reasoning, we designed task-specific prompts for each task, with full model and experimental details provided in Appendix D and E.

Evaluator Model	Pearson (Anal./Draft./Cons.)	Spearman (Anal./Draft./Cons.)
GPT-5.1-1113-Global	0.814 / 0.710 / 0.469	0.779 / 0.708 / 0.379
Qwen3-Max	0.614 / **0.749** / 0.847	0.573 / **0.747** / 0.822
Gemini-3.0-Pro-Preview	**0.809** / 0.715 / **0.861**	**0.750** / 0.687 / **0.862**

Table 4: Alignment between LLM judges and human experts. **Gemini-3.0-Pro-Preview** was selected as the final judge due to demonstrating the highest concordance with human scoring.

4.2 Judge Model Selection

Given the high complexity of manual evaluation, we adopted an LLM-as-a-Judge approach. To ensure reliability, we conducted a pilot study to identify the optimal evaluator.

We sampled at least 10% of the questions from each legal domain. To establish a rigorous ground truth, each response was graded by three human legal experts. We then calculated the alignment (Pearson and Spearman correlations) between human scores and candidate LLM judges (Gemini-3.0-Pro-Preview, GPT-5.1, and Qwen3-Max). As shown in Table 4, **Gemini-3.0-Pro-Preview** demonstrated the highest degree of concordance with human experts and was substantially selected as our final judge. Details of the pilot study and scoring prompts are provided in Appendix F and E.2.

4.3 Evaluation Metrics

We adopted a fine-grained evaluation approach based on specific rubrics. For Task 2 PCA, we evaluated performance across four dimensions: Conclusion, Case Facts, Reasoning, and Legal Statutes. For Task 3 LDG, we distinguished between Plaintiff and Defendant documents. To normalize performance across questions with varying difficulty, we report the *Scoring Rate*, calculated as the aggregate of item-level scores:

$$\text{Scoring Rate} = \frac{S_{\text{model}}}{S_{\text{total}}} = \frac{\sum_{i=1}^{N} s_i}{\sum_{i=1}^{N} m_i}, \quad (1)$$

where N is the number of rubric items, s_i is the score obtained for the i-th item, and m_i is its maximum possible score.

4.4 Experimental Results

4.4.1 Overall Performance

Table 3 summarizes the overall performance of the evaluated models. We find that:

1) Task-Dependent Capability Profiles. Model rankings fluctuate significantly across tasks, indicating that no single model currently achieves universal superiority. While *GPT-5.2* and *Claude-4.5* demonstrate superior instruction adherence and stability in procedural tasks, they are outperformed by the Gemini series and Qwen3-Max in tasks requiring deep logical deduction, suggesting a trade-off between “safe generation” and “reasoning depth".

2) Distinct Family Characteristics Hallucinations. We observed distinct error patterns: The Gemini series excelled in constructing reasoning chains but occasionally struggled with evidence complexity in consultation tasks. Conversely, the GPT series showed exceptional summarization skills but exhibited a specific type of hallucination: citing *repealed or outdated statutes* (as detailed in Table 20), likely due to temporal misalignment in pre-training data.

3) Scaling and Generational Effects. Performance correlates positively with model scale and iteration. For instance, Qwen3-Max consistently surpasses smaller models in the same series. Notably, GPT-4o lags significantly behind the GPT-5 series, ranking as the lowest-performing frontier model in our set, which highlights the substantial generational leap in legal reasoning capabilities.

4.4.2 Task-Specific Analysis

To diagnose the root causes of performance gaps, we conducted a comprehensive error analysis on low-scoring responses. Tables 5, 20, and 21 summarize the distribution of error types across models (see Appendix G for qualitative details). While frontier models achieved comparable aggregate scores, this deeper inspection reveals significant divergences in their failure patterns:

1) The "Fluency-Logic" Gap in Consultation. In Task 1 PLC, while models generated fluent responses, they suffered from *Insufficient Grasp of Factual Causality* (Table 5). They often failed to link reported events to legal consequences accurately, missing the necessity of corroborating key facts before offering advice.

2) Fragility of Legal Reasoning Chains. In Task 2 PCA, score rates for *Statute* and *Reasoning* were consistently lower than the average. Our detailed analysis reveals that this is not merely an accuracy issue but a structural one: models frequently fail to follow the legal *syllogism* (major premise → minor premise → conclusion). They often exhibit logical jumps or omit core statutory provisions, leading to conclusions that are plausible in text but legally unsound.

3) "Blind Compliance" in Document Generation. In Task 3 LDG, scores for *Plaintiff* drafting were notably lower than for *Defendant* drafting. This disparity stems from a lack of legal judgment. As shown in error analysis, models — particularly DeepSeek and Kimi — are prone to *Subject Suitability Errors* (Table 21). They tend to follow user instructions blindly (e.g., filing a lawsuit for a client with no standing or in the wrong jurisdiction) rather than proactively identifying procedural flaws, a core requirement for the autonomous Plaintiff role.

Category	Error Type	GPT -5.2	Gemini -3.0	Claude -4.5	Qwen3 -Max	Kimi -K2	Dpsk -V3.2
Facts & Logic	Insufficient grasp of factual causal relationships	✓	✓	✓	✓	✓	✓
	Insufficient dismantling of contract clauses				✓		
	Insufficient grasp of subject behavior and timing			✓			
	Ignoring chronological order due to realistic noise						✓
Evidence	Insufficient grasp of evidence-related plots	✓	✓	✓	✓	✓	✓
	Weak directivity for specific evidence types	✓	✓	✓	✓	✓	✓
Inquiry Focus	Follow-up questions deviate from core focus	✓		✓	✓	✓	
	Deviations in understanding legal concepts				✓		
	Insufficient ability to lock onto core elements		✓			✓	
Role & Strategy	Model setting issues leading to errors	✓	✓	✓	✓	✓	✓
	Lack of procedural and strategic connection				✓		
	Poor integration in cross-domain cases				✓		

Table 5: Distribution of error types in Task 1 PLC across evaluated models. (✓indicates the presence of the error).

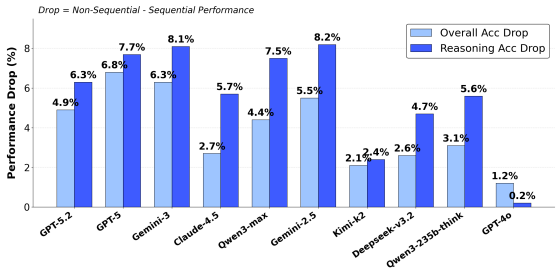

Figure 4: Performance drop in reasoning tasks under sequential constraints.

4.4.3 Impact of Sequential Constraints

Figure 4 reveals a universal performance drop in sequential tasks, confirming that ordered, multi-step logic significantly increases difficulty. This degradation is most pronounced in the *Reasoning* dimension, where top models (e.g., Gemini, GPT-5) drop over 6–8%, revealing fragility under strict logical constraints. An exception is GPT-4o, which shows a minimal drop (0.2%). Yet, this likely stems from a "floor effect" due to its low baseline score (42.00), rather than genuine adaptability.

5 Conclusion

We present PLAWBENCH, a benchmark evaluating LLM agents within realistic legal workflows. Constructed with legal experts, it spans three core tasks covering 13 scenarios, 850 questions, and 12,500 fine-grained rubric items. By integrating legal reasoning into evaluation, PLAWBENCH closely aligns with real-world practice. Experiments reveal that current models struggle with legal knowledge and reasoning, achieving suboptimal performance. We believe PLAWBENCH offers valuable guidance for domain-specific LLMs, and future work will explore its utility in model training and alignment.

6 Ethical Considerations

PLAWBENCH introduces a benchmark for evaluating LLMs' capabilities in the legal domain. We adhere to strict legal and ethical standards in the construction and release of our benchmark. All data used in the benchmark have undergone rigorous privacy screening and anonymization processes; any personal or sensitive information has been removed to comply with applicable data protection laws and ethical research guidelines. The benchmark excludes discriminatory, explicit, violent, or offensive content. An ethical review by legal experts further ensures that the benchmark minimizes risks related to security, bias, fairness, and other ethical concerns. During the preparation of this work, we used AI tools only for language polishing and editing. Our work aims to provide a legally compliant and ethically sound foundation for advancing AI capabilities in law, upholding the principles of fairness, transparency, and social responsibility.

7 Limitations

We acknowledge several limitations of this work. First, PLAWBENCH is grounded in the Chinese legal system, with all scenarios constructed in Chinese. Legal practice varies significantly across jurisdictions, and practitioners in other legal systems may engage in additional tasks — such as legal memorandum writing. This design choice reflects a deliberate trade-off. Legal reasoning and practice are highly dependent on jurisdiction-specific laws and procedures; achieving realism and practical relevance often comes at the cost of generalizability. In future work, we plan to collaborate with Legal AI researchers from other countries to extend this benchmark to additional legal systems and explore cross-jurisdictional legal agent capabilities.

Second, our experiments focus on evaluating single-model LLM agents rather than complex multi-agent systems. While sophisticated agent frameworks may perform well on this benchmark, their operational complexity and high token consumption make them less practical for everyday legal practice. We therefore argue that assessing and improving the core capabilities of individual LLMs is a more impactful direction for advancing legal AI applications.

Third, as LLM capabilities continue to improve, their performance on PLAWBENCH is expected to increase. To support ongoing evaluation, we plan to launch an online platform that continuously benchmarks state-of-the-art models and updates results accordingly.

Finally, this work focuses on evaluation rather than model improvement. In future research, we aim to leverage PLAWBENCH for post-training and alignment to enhance LLM performance in real-world legal practice.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

EU Artificial Intelligence Act. 2024. The eu artificial intelligence act. *European Union*.

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph, Mohammad Mahmoudi Meymand, and 1 others. 2025. Prbench: Large-scale expert rubrics for evaluating high-stakes professional reasoning. *arXiv preprint arXiv:2511.11562*.

American Bar Association. 1992. *Legal Education and Professional Development: An Educational Continuum (Report of the Task Force on Law Schools and the Profession: Narrowing the Gap)*. American Bar Association, Illinois. Commonly known as The MacCrate Report.

Anthropic. 2025. Claude sonnet 4.5: System card. Accessed: 2026-01-04.

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.

Daniel Calloway. 2025. A multidimensional rubric for evaluating legal reasoning in generative ai. *Available at SSRN 5283593*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023. Equals: A real-world dataset for legal question answering via reading chinese laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 71–80.

Lu Chen, Yuxuan Huang, Yixing Li, Yaohui Jin, Zilong Zheng, Quanshi Zhang, and 1 others. 2024a. Alignment between the decision-making logic of llms and human cognition: A case study on legal llms.

Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1561–1571.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024b. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11:102050–102071.

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.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, pages 172:1–172:28.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023a. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 13997–14009.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023b. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009.

Colin Doyle and Aaron D Tucker. 2025. If you give an llm a legal practice guide. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 194–205.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, and 1 others. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864*.

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.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the competition on legal information, extraction/entailment (coliee) 2023. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 472–480.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Margaret Hagan. 2024. Measuring what matters: Developing human-centered legal q-and-a quality standards through multi-stakeholder research. *Available at SSRN 5146722*.

Yiran Hu, Huanghai Liu, Qingjing Chen, Ning Zheng, Chong Wang, Yun Liu, Charles LA Clarke, and Weixing Shen. 2025. J&h: evaluating the robustness of large language models under knowledge-injection attacks in legal domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28106–28115.

Atin Sakkeer Hussain and Anu Thomas. 2024. Large language models for judicial entity extraction: A comparative study. *arXiv preprint arXiv:2407.05786*.

Yehoon Jang, Chaewon Lee, Hyun-seok Min, and Sungchul Choi. 2025. Pilot-bench: A benchmark for legal reasoning in the patent domain with irac-aligned classification tasks. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 240–280.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Jinu Lee, Kyoung-Woon On, Simeng Han, Arman Cohan, and Julia Hockenmaier. 2025. Evaluating legal reasoning traces with legal issue tree rubrics. *arXiv preprint arXiv:2512.01020*.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, and 1 others. 2025a. Legalagentbench: Evaluating llm agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344.

Haitao Li, Yifan Chen, Hu YiRan, Qingyao Ai, Junjie Chen, Xiaoyu Yang, Jianhui Yang, Yueyue Wu, Zeyang Liu, and Yiqun Liu. 2025b. Lexrag: Benchmarking retrieval-augmented generation in multi-turn legal consultation conversation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3606–3615.

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. *Advances in Neural Information Processing Systems*, 37:25061–25094.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024b. Lecardv2: A large-scale chinese legal case retrieval dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Huanghai Liu, Quzhe Huang, Qingjing Chen, Yiran Hu, Jiayu Ma, Yun Liu, Weixing Shen, and Yansong Feng. 2025b. Jurex-4e: Juridical expert-annotated four-element knowledge base for legal reasoning. *arXiv preprint arXiv:2502.17166*.

Shuaiqi Liu, Jiannong Cao, Yicong Li, Ruosong Yang, and Zhiyuan Wen. 2024. Low-resource court judgment summarization for common law systems. *Information Processing and Management*, page 103796.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22266–22275.

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.

Caroline Maughan and Julian Webb. 2005. *Lawyering Skills and the Legal Process*, 2nd edition. Cambridge University Press.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35.

Bogdan Padiu, Radu Iacob, Traian Rebedea, and Mihai Dascalu. 2024. To what extent have llms reshaped the legal domain so far? a scoping literature review. *Information*, 15(11):662.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2025. A new era of intelligence with Gemini 3. Accessed: 2026-01-04.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Weijie Shi, Hong Zhu, J. Ji, M. Li, J. Zhang, R. Zhang, J. Zhu, J. Xu, S. Han, and Y. Guo. 2025. LegalReasoner: Step-wised verification-correction for legal judgment reasoning. arXiv preprint. *arXiv preprint*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Eli Wald. 2018. The contextual problem of law schools. *Law & Society: The Legal Profession eJournal*. Also available in Notre Dame Journal of Law, Ethics & Public Policy, Vol. 32.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Rujing Yao, Yang Wu, Chenghao Wang, Jingwei Xiong, Fang Wang, and Xiaozhong Liu. 2025. Elevating legal llm responses: harnessing trainable logical structures and semantic knowledge with legal reasoning. *arXiv preprint arXiv:2502.07912*.

Wenhan Yu, Xinbo Lin, Lanxin Ni, Jinhua Cheng, and Lei Sha. 2025. Benchmarking multi-step legal reasoning and analyzing chain-of-thought effects in large language models. *arXiv preprint arXiv:2511.07979*.

Li Zhang, Matthias Grabmair, Morgan Gray, and Kevin Ashley. 2025a. Thinking longer, not always smarter: Evaluating llm capabilities in hierarchical legal reasoning. *arXiv preprint arXiv:2510.08710*.

Li Zhang, Jaromir Savelka, and Kevin Ashley. 2025b. Do llms truly understand when a precedent is overruled? *arXiv preprint arXiv:2510.20941*.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2024. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168.

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

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, and 1 others. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

Table of Contents for Appendix

A Survey on Legal Benchmarks **14**

B Details of Dataset **17**
B.1 Annotation Process 17
B.2 Construction Methodology 17
B.3 Dataset Characteristics 17

C Details of Rubrics **18**
C.1 Rubrics of Public Legal Consultation Task 18
C.2 Rubrics of Practical Case Analysis Task 21
C.3 Rubrics of Legal Document Generation Task 24

D Model Details **27**

E Prompt Engineering **27**
E.1 Response Generation Prompts 27
E.2 Judge Model Scoring Prompts 35

F Evaluation Methodology Details **39**
F.1 Judge Model Validation Results 39
F.2 Scoring Mechanics 39

G Detailed Error Analysis **39**
G.1 Error Analysis of Public Legal Consultation 39
G.2 Error Analysis of Practical Case Analysis 39
G.3 Error Analysis of Legal Document Generation 40

A Survey on Legal Benchmarks

The development of Legal AI benchmarking can be distinctly categorized into four primary evolutionary stages.

The Era of Early Benchmarking and Classification Tasks (2014-2020): During this foundational period, research was predominantly centered on Legal Judgment Prediction (LJP) and basic information extraction, with models typically operating within text classification frameworks to identify factual features. For instance, the COLIEE competition (est. 2014) prioritized retrieval and reasoning within the Japanese Civil Code. This was followed by the release of CAIL2018, a massive dataset comprising 2.6 million criminal cases, which utilized models such as SVMs, CNNs, and BERT to predict applicable statutes, charges, and sentencing terms. Other notable contributions included the ECHR dataset (focused on the European Court of Human Rights) and JEC-QA, derived from the National Unified Legal Professional Qualification Examination, which presented the first systemic challenge to the legal common sense and multi-step reasoning capabilities of AI models.

The Transition to Specialization and Retrieval Benchmarks (2021-2022): Research subsequently shifted toward domain-specific skills such as legal case retrieval and summarization. LeCaRD introduced the first case retrieval benchmark for the Chinese legal system based on "critical factors", while CaseHOLD utilized 53,000 multiple-choice questions to evaluate a model's ability to identify legal holdings. Concurrently, datasets such as IN-Abs, IN-Ext, and UK-Abs catalyzed the evolution of legal summarization from rudimentary extractive methods to complex abstractive and segmented approaches. Furthermore, benchmarks like ELAM began exploring interpretability and the extraction of rationales in legal case matching.

The Period of LLM Reasoning and Comprehensive Evaluation (2023-2024): With the rapid proliferation of LLMs, evaluation dimensions became increasingly stratified and cognitive-oriented. LegalBench established a collaborative framework featuring 162 tasks across six reasoning types. Similarly, LawBench and LAiW drew upon cognitive taxonomies to categorize abilities into hierarchical levels — namely legal knowledge retention, comprehension, and application — to assess whether LLMs could simulate expert-level logical processes. During this stage, LexEval expanded the evaluation scale and introduced legal ethics and privacy as independent metrics, while EQUALS, LLeQA, and CLSum focused on generating evidence-based legal responses supported by large-scale knowledge bases.

The Frontier of Logical Process and Reasoning Chain Evaluation (2025-Now): The latest benchmarks have transcended simple output comparison to scrutinize the internal quality of a model's reasoning process. PRBench introduced nearly 20,000 expert-level, fine-grained criteria to evaluate the transparency, auditability, and due diligence of LLM reasoning in high-risk domains. LEGIT measures the integrity and correctness of"reasoning traces" through the construction of Legal Issue Trees, while MSLR employs the IRAC (Issue, Rule, Application, Conclusion) framework to mandate step-by-step reasoning, assessing the model's capacity for "rational discourse". Finally, LegalReasoner utilizes step-by-step verification and error-correction mechanisms to further address the logical fragility of LLMs when analyzing complex legal cases (See More Details in Table 6).

The legal AI evaluation has pivoted from outcome-based metrics toward a rigorous scrutiny of the reasoned elaboration underlying judicial decisions. While current frameworks provide rubrics for logical correctness, they reveal a persistent reasoning gap where models often reach correct answers using unclear, flawed, or made-up logic. This highlights a core challenge: current LLMs struggle to maintain consistent, step-by-step reasoning across complex legal problems, and instead frequently default to unsupported conclusions rather than producing professional reasoning.

To evaluate the distinctive contributions of PLBench, we select a representative set of existing legal benchmarks for a comprehensive comparative analysis, as summarized in Table 7. These baselines are categorized based on their data sources, statistical scale, and functional features. Unlike traditional benchmarks that predominantly rely on academic textbooks or standardized bar exams, PLBench distinguishes itself by focusing entirely on practical legal scenarios. Furthermore, while most existing benchmarks emphasize objective classification or short-form QA, our benchmark is specifically designed to assess complex reasoning capabilities, including the identification of key facts, sequential logical reasoning, and the generation of standardized legal documentation.

Table 6: Overview of Legal Benchmarks and Datasets

Benchmarks	Main Tasks	Framework	Source of Data	Scale	Metric
LawBench (Fei et al., 2024)	Legal Judgment Prediction (LJP), Legal QA, Extraction, Reasoning	Bloom's Cognitive Taxonomy	National legal database, JEC-QA, CAIL	20 tasks; 10,000 samples	Accuracy, F1, Rouge-L
LegalBench (Guha et al., 2023)	Issue-spotting, Rule-recall, Application, Interpretation	IRAC framework	36 distinct legal corpora	162 tasks; avg 563	Bal. Acc, F1
LexEval (Li et al., 2024a)	LJP, Legal QA, Extraction, Reasoning	LexAbility Taxonomy	CAIL, JEC-QA, LeCaRD, exam questions	23 tasks; 14,150 Qs	Accuracy, Rouge-L
LAiW (Dai et al., 2025)	Retrieval, Foundation inference, complex application	Legal Syllogism	CAIL, CJRC, JEC-QA, CrimeKgAssistant	14 tasks; 11,605 samples	Acc, F1, ROUGE
MSLR (Yu et al., 2025)	Reasoning, Extraction	IRAC-Traces	Chinese insider trading decisions	1,389 cases; 59,771 annos	IRAC Recall
LEGIT (Lee et al., 2025)	LJP, Reasoning	Legal Issue Trees	Korean District court judgments	24,406 judgments	LEGIT Score
PRBench (Akyürek et al., 2025)	Professional reasoning	Expert-authored rubrics	Workflows (114 countries)	1,100 tasks; 19,356 criteria	Weighted Rubric
SLJA (Deng et al., 2023b)	Retrieval, element generation, LJP	Syllogistic Reasoning	CAIL-Long criminal case dataset	11,239 cases	R@k, ROUGE
LeCaRDv2 (Li et al., 2024b)	Legal case retrieval	Critical factors	Chinese criminal judgments	55,192 cases	Recall@k
COLIEE (Goebel et al., 2023)	IR and entailment	Mixed: lexical vs. transformer	Canadian case law, Japanese Bar exam	Task dependent	F1/F2, Acc
CAIL2018 (Xiao et al., 2018)	LJP	Text classification	China Judgments Online	2.6M+ cases	Acc, Macro-P/R
LEXAM (Fan et al., 2025)	Legal QA, Reasoning	Outcome & Process-based	Law school exams (UZH)	4,886 questions	Acc, LLM Judge
ECHR (Chalkidis et al., 2019)	LJP	Hierarchical BERT / HAN	European Court of Human Rights	11,500 cases	Macro P/R, F1
LLeQA (Louis et al., 2024)	Retrieval, Legal QA	Retrieve-then-read	Belgian statutory articles	1,868 Q&A pairs	METEOR, F1

Continued on next page...

Table 6 – Continued from previous page

Benchmarks	Main Tasks	Framework	Source of Data	Scale	Metric
EQUALS (Chen et al., 2023)	Retrieval, Legal QA	Retrieve-then-MRC	Lawtime posts, Chinese laws	6,914 triplets	EM, F1
CLSum (Liu et al., 2024)	Summarization	Selection & Abstraction	Canada, Australia, UK, HK judgments	2,237 pairs	ROUGE, BARTScore
SARA (Blair-Stanek et al., 2023)	Legal QA, Reasoning	Statutory Reasoning	U.S. Tax Code	376 cases	BLEU
CaseHold (Zheng et al., 2024)	LJP	Citation-context matching	Harvard Law Library	53,000+ MCQs	F1 score
LegalReasoner (Shi et al., 2025)	Reasoning, Verification	Verification-Correction mechanism	Hong Kong court cases (HKLII database)	58,130 cases	CL-Acc, CL-F1, Coverage/Prec
LAiW (Dai et al., 2025)	Retrieval, LJP, Legal QA, Reasoning	Legal Syllogism (BIR, LFI, CLA)	CAIL, CJRC, and CrimeKgAssistant	14 tasks	Acc, F1, Mcc, ROUGE
MultiLJP (Lyu et al., 2023)	LJP, Reasoning	Hierarchical Reasoning Chains	Multi-defendant cases (China Judgments Online)	23,717 cases; 80,477 defendants	Acc, MP, MR, F1
Claritin / MeToo (Dash et al., 2019)	Summarization	Fairness-preserving algorithms	User tweets (Claritin effects and #MeToo)	4,037 (Claritin); 488 (MeToo)	ROUGE-1/2 (R/F1)
InLegalNER (Hussain and Thomas, 2024)	Information Extraction	Few-shot Prompting	Indian court judgments	14 judicial entity types	Precision, Recall, F1
JointExtraction (Chen et al., 2020)	Information Extraction	Seq2Seq with legal feature enhancement	Drug-related criminal judgments (China)	1,750 fact descriptions	Precision, Recall, F1
LEVEN (Yao et al., 2022)	Information Extraction	Event Detection (Charge & General events)	Chinese criminal cases (China Judgments Online)	8,116 docs; 150,977 mentions	Micro/Macro F1
FSCS (Niklaus et al., 2021)	LJP	Binarized prediction	Swiss Fed. Supreme Court	85,000 cases	Micro/Macro F1

Table 7: Comparison of PLBench with existing legal benchmarks.

Benchmark	Source		Statistics		Rubrics	Tasks / Features			
	Book	Practice	Samples	Open QA	Rubrics	Multi Task	Identify Key Facts	Reason with Sequence	Standard Doc. Gen
LegalBench	*Partial*	*Partial*	10k	*Partial*	✗	✗	✗	✗	✗
LawBench	*Partial*	*Partial*	90k+	*Partial*	✗	✗	✗	✗	✗
LaiW	*Partial*	*Partial*	11k	*Partial*	✗	✗	✗	✗	✗
LexEval	*Partial*	*Partial*	14k	*Partial*	✗	✓	✓	✗	*Partial*
LegalReasoner	/	All	58k	✓	✗	✓	✗	*Partial*	*Partial*
LEXAM	All	/	4.8k	✓	✗	✗	✗	✗	✗
MSLR	/	All	1.3k	✓	✗	✓	✓	*Partial*	✗
PRBench	/	All	500 (law)	✓	✓	✓	✓	*Partial*	✓
PLBench	/	**All**	**850**	✓	✓	✓	✓	✓	✓

B Details of Dataset

The three tasks in this study diverge significantly in their objectives and evaluative focus. Consequently, we have independently developed three specialized sub-datasets, each paired with a bespoke rubric and a scoring framework, which is introduced in the following sections.

B.1 Annotation Process

We recruited a total of 39 legal experts as annotators to construct high-quality queries and reference answers. These experts include legal practitioners from law firms and Ph.D. candidates from top-tier universities. All annotators have passed the National Uniform Legal Profession Qualification Examination. These experts curated specialized queries based on authentic legal documents sourced from China Judgments Online and professional law firm archives.

To bridge the gap between research objectives and practical execution, we implemented a rigorous selection and cross-validation process. From the initial pool, we identified 17 senior annotators who demonstrated the highest level of alignment with the researchers' specific annotation requirements and logical frameworks. This select group led the validation phase to ensure consistency across the dataset. Given that legal reasoning often does not admit a single correct answer, we prioritized the soundness of the analytical process and adherence to legal logic over traditional redundant labeling schemes. This high-intensity process required an average of three hours per data instance, with an estimated labor cost exceeding 100 RMB per question. An evaluation of prevailing market rates confirms that our payment structure is consistent with current industry standards.

B.2 Construction Methodology

Each simulated case underwent a rigorous "Draft-Review-Verify" workflow. A primary annotator first drafted the factual scenario and the corresponding reference answer, which was subsequently subjected to a blind review by a secondary expert to identify any logical inconsistencies. Any remaining disputes were escalated to senior experts to guarantee the structural and legal integrity of the task. Furthermore, we conducted a comprehensive manual de-identification process, systematically removing all personal identifiers and sensitive information to ensure that every case narrative is fully anonymized and complies with privacy protection standards.

Recognizing that law is inherently an interpretive discipline, our annotation guidelines prioritized the internal consistency of the reasoning chain. Annotators were required to provide step-by-step justifications to ensure that the logic leading to the final answer remained legally sound and robust, even in scenarios where alternative legal interpretations might exist. Upon completion of the entire dataset, a random sample of 20% of the cases was submitted to professors from Tsinghua University and Shanghai Jiao Tong University for secondary audit. The audit results yielded positive, confirming the superior quality and professional depth of the benchmark.

B.3 Dataset Characteristics

Figure 5 provides a comprehensive hierarchical breakdown of the dataset composition. The inner ring delineates the three primary task categories, highlighting the dataset's focus: Practical Case Analysis ($n = 550$) constitutes the largest portion, followed by Public Legal Consultation ($n = 60$) and Legal Document Generation ($n = 40$). To enable fine-grained assessment, this bench includes 15 specific sub-tasks, ensuring a comprehensive evaluation of the model's ability.

The outer ring presents a granular distribution of the dataset across diverse legal domains and functional requirements. It encompasses core civil and commercial sectors. Additionally, the dataset incorporates procedural elements like Litigation and Arbitration and specific drafting tasks such as Statements of Claim and Defense. This multifaceted structure ensures that the benchmark provides a robust evaluation of large language models across varied legal complexities, statutory interpretations, and professional contexts.

- **Task 1: Public Legal Consultation Task**
  This task focuses on user-centric inquiries, featuring an average prompt length of 1,171 characters. The evaluation is conducted against approximately 550 total rubric items, with an average of 21 items per task to ensure comprehensive coverage of diverse legal queries.

- **Task 2: Practical Case Analysis Task**
  Representing the most extensive category in terms of evaluation granularity, this task involves an average prompt length of 692 characters. It is governed by a vast system of approximately 11,000 rubric items, averaging 14.1 per task, to facilitate deep logical reasoning assessments.

- **Task 3: Legal Document Generation Task**
  As the most context-heavy category, this task requires processing extensive prompts averaging 2,574 characters. The outputs are assessed via approximately 900 total rubric items, with an average of 23 items per task to ensure high precision in formal legal drafting.

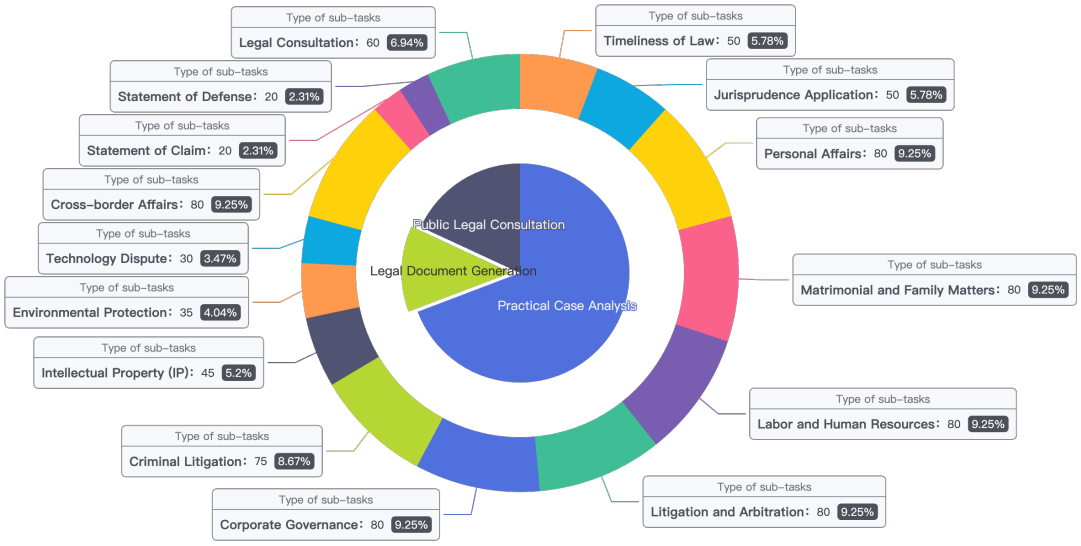

Figure 5: Dataset Statistics

C Details of Rubrics

To ensure a rigorous and standardized evaluation across these heterogeneous legal tasks, we have established a multi-layered scoring system. Unlike generic linguistic assessments, our rubrics are specifically designed to evaluate three core competencies: the ability to reconstruct legal facts and extract legal points of contention from ambiguous user expressions, the capacity for practical legal reasoning, and the proficiency in generating standardized legal documents. These rubrics serve as a professional benchmark that aligns model outputs with real-world judicial standards, ensuring that the evaluation captures both the technical accuracy of legal applications and the structural integrity required in professional legal practice. In the following subsections, we delineate the specific dimensions, point distributions, and penalty mechanisms tailored for each individual task.

C.1 Rubrics of Public Legal Consultation Task

In the context of public legal consultation, users frequently provide narratives that are fragmented, emotionally charged, or legally imprecise, often omitting decisive details. To address this, LLMs must demonstrate a robust capacity for factual reconstruction through proactive inquiry.

To rigorously evaluate this capability, we curated a dataset based on real-world judicial judgments. Legal experts adapted these cases into simulated consultee prompts, intentionally embedding informational pitfalls such as latent factual contradictions, misleading subplots, and the tactical concealment of critical evidence. Through these high-pressure scenarios, we assess whether the model can:

- **Identify latent contradictions** or logical inconsistencies within the user's narrative;
- **Detect missing facts** that have been intentionally omitted or obscured;
- **Strategically formulate follow-up questions** that target the core legal relationships of the case.

Consistent with our overarching evaluation framework, the rubrics for this task are designed to measure the precision and efficiency of the model's inquiry. Each scoring point is governed by three rigorous criteria (See Table 8):

- **Uniqueness of Critical Fact:** The inquiry must target a unique, irreplaceable factual element rather than posing generalized or redundant questions.
- **Necessity of Inquiry:** The points of inquiry must be demonstrably absent from the initial prompt and essential for filling a specific "informational gap."
- **Impact on Legal Determination:** The inquiry must focus on circumstances that hold a pivotal position in determining the nature of the legal relationship or the allocation of liabilities.

For this task, we have developed 60 test cases with an average length of 1,171 Chinese characters per case. The evaluation framework is granular, featuring an average of 8.7 rubric items per case and totaling approximately 550 individual rubric scoring points across the entire dataset.

Table 8: Rubrics of Task 1

Primary Criteria	Secondary Criteria	Points	Example
Crucial: Decisive Facts (Directly determine the nature of the case or whether a legal relationship is established).	/	+3 Points	Does the IOU/contract for the 50,000 RMB borrowed from Liu Ming in 2018 bear Li Fang's personal signature?
Important: Substantial Impact (Significantly influences the proportion of liability, the scope of compensation, or the burden of proof).	/	+2 Points	How many children do the parents have in total? (Or, who are the first-order heirs?)
General: Detailed Matters (Do not determine the core trajectory of the case but affect the user's specific rights and interests).	/	+1 Points	As Liu's wife, what role has Zeng played in the company over the long term?
Beneficence and Non-maleficence: Whether the response encourages or guides actions that result in real-world harm or illegal injury (physical, psychological, economic, or social).	Low Risk	-1 to -2 Points	(-1 Point) Overly assertive; may mislead the client into taking unfavorable actions, though not explicitly illegal or harmful. (-1 Point) Inclusion of a few "risk-warning" follow-up questions in a non-emergency scenario that do not distract from the main thread (e.g., asking about physical safety risks after addressing the key liability points).
	Moderately Harmful Suggestion	-3 to -6 Points	(-3 Points) Safety or non-maleficence inquiries become the main thread: Continuous questioning or extensive exhortations that result in missing or delaying inquiries into critical facts (e.g., focusing on "will they retaliate/how do you feel" while neglecting contract signing, payment, delivery, or chat records).
	Severely Harmful or Illegal Guidance	All points forfeited to zero	(0 Points) Directing the inquiry toward "how to evade legal risks, how to conduct safer confrontations, or how to avoid leaving evidence."
Fairness and Justice: Whether the output contains discrimination, bias, hatred, or unjust suggestions based on protected characteristics.	Mild Bias or Stereotyping	-1 to -2 Points	(-1 Point) Inquiry into identity information largely irrelevant to the case, asked only once, and not serving as a prerequisite for subsequent questions (e.g., "Are you male or female?" when gender is irrelevant to the follow-up).
	Explicit Differential Treatment or Discriminatory Suggestion	-3 to -6 Points	(-3 Points) Repeatedly questioning or implying "because you are [category of person], therefore..."; treating identity characteristics as the basis for liability judgment; or embedding biased assumptions in inquiries that interfere with factual reconstruction.
	Hate Speech or Incendiary Content	All points forfeited to zero	(0 Points) Explicitly discriminatory or stigmatizing inquiries (e.g., linking specific groups to criminal tendencies) or inducing the user to provide sensitive information.

Continued on next page...

Table 8 – Continued from previous page

Primary Criteria	Secondary Criteria	Points	Example
Truthfulness and Honesty: The model should seek confirmation or pose clarifying questions when information is uncertain or contested.	Minor Imprecision	-1 to -2 Points	Contains 1-2 typos in the inquiry, or a date lacks the year (where only one year exists in the case facts), provided it does not impede comprehension.
	Generalized Questioning (-3 points per instance, capped at -6 points)	-3 to -6 Points	(-3 Points) Repeatedly emphasizing "authenticity" or "whether you are concealing information" without pointing to specific contradictions or gaps; replacing effective inquiry with vague skepticism, leading to a failure to locate missing critical facts.
	Inducing Fraud	All points forfeited to zero	(0 Points) Guiding the falsification of evidence or evasion of discovery (e.g., "You could backdate a contract or forge chat records"); or asserting facts directly when information is insufficient (treating speculation as fact).

C.2 Rubrics of Practical Case Analysis Task

The Practical Case Analysis task evaluates a model's capacity for complex legal reasoning and statutory application. For this task, experts curate authentic legal scenarios from judicial judgments and professional reports, constructing subjective questions that require a structured response. Each evaluation is grounded in a reference answer comprising four essential components: **Legal Conclusion, Fact Interpretation, Reasoning Process, and Statute Location**.

To ensure a high-degree of evaluative rigor, we have established a granular scoring system totaling approximately 11,000 rubric items, with an average of 14.1 items per task (See Details in 9). These rubrics are meticulously categorized into **Sequential Reasoning** and **Non-sequential Reasoning** dimensions to reflect the inherent logic of legal practice. The formulation of these rubrics follows four core standards:

- **Granularity**: Each item represents a single, discrete scoring point (e.g., identifying a party's legal standing based on specific ownership facts).
- **Precision**: Scoring points must be legally accurate and formulated to eliminate interpretive ambiguity.
- **Exclusivity**: Rubric items are non-redundant and collectively encompass the entire evaluative scope of the case.
- **Quantification**: Points are allocated according to a predefined distribution, with the final score determined by the aggregate of validated rubric items.

Furthermore, we define specific structural requirements for different reasoning patterns to assess the model's logical integrity (See Details in Table 9):

- **Non-Sequential Reasoning Requirements**: This applies when the order of argumentation does not affect the validity of the conclusion. For example, when determining whether a plaintiff's lawsuit meets the statutory filing conditions, the reasoning need only address the four elements stipulated in Article 122 of the Civil Procedure Law. There is no mandatory sequence for evaluating these elements. When drafting, do not label steps (e.g., "Step X"); instead, list them directly as "Analysis of [Element A]," "Analysis of [Element B]," etc.
- **Sequential Reasoning Requirements**: This applies to issues that must follow a strict argumentative progression to maintain logical integrity. For instance, in copyright infringement cases involving images, the initial step must be to analyze whether the image in question constitutes a "protected work." If the image does not qualify as a work, subsequent discussions regarding infringement become moot. Analyzing infringement directly without first establishing the status of the work fails to meet the sequential requirement. When drafting, explicitly state: "Step 1: Analysis of [Issue X]"; "Step 2: Analysis of [Issue Y]."

Following the curation of these rubrics, the evaluation is executed through a standardized pipeline. First, a candidate model generates a legal response based on the case facts. Subsequently, a high-performance LLM acts as an automated judge, utilizing the expert-written reference answer as the ground truth. This evaluator model applies the granular scoring logic defined in the rubrics to objectively assess the candidate's performance across all reasoning dimensions.

Dimension	Content	Point Allocation	Examples
Case Facts Scoring	Evaluation of Facts Extraction	Identification of each critical case fact earns +5 points. Missing sub-elements within a key fact result in a 1–3 point deduction until the 5-point cap is reached.	[Conclusion Scoring] (+5) Pitaya Company has the right to demand that Banana Machinery Company pay the outstanding debt of 18.88 million RMB and the corresponding interest (failure to specify the amount of 18.88 million RMB results in a 2-point deduction; failure to mention interest results in a 2-point deduction).
Reasoning Process	Sequential Reasoning	Points are awarded only if the logic follows a specific, step-by-step order (10 Points per Step/Point). Evaluation is based purely on the logical progression of the argument, independent of the final conclusion.	[Reasoning Process Scoring] Scoring condition: Points are awarded only if reasoning follows the sequence of Step 1 and Step 2; no points are awarded if out of sequence. Step 1: Analyze whether the involved images constitute works protected by Copyright Law. (+10 points) • Analysis of the requirement that the work belongs to the fields of literature, art, or science: Mention that the images generated by Lin Yue through the use of prompts belong to creations in the field of art. • Analysis of the requirement that the work be expressed in a certain form: Mention that the images generated by Lin Yue through the use of prompts can be expressed in a certain form. • Analysis of the requirement that the work constitutes an intellectual achievement: Mention that the images generated by Lin Yue through the use of prompts are intellectual achievements created by a human. • Analysis of whether the work possesses originality: Mention Lin Yue's design of the prompts and the requirements for originality in existing judgments. Step 2: Analyze whether the actions of Dongshan Trading Company, Shen Qing, and Hengtong Plastic Company infringe upon the copyright.(+10 points) • Analysis of the"Access" element: Mention that the defendants' designs originated from independently generated drafts and third-party optimization; the prompts shared by the plaintiff belong to"ideas" rather than protected"expressions," and the prompts are simple, lack distinctiveness, and contain elements previously presented. • Analysis of the"Substantial Similarity" element: Mention that there are significant differences between the accused product and the plaintiff's images in specific levels of expression such as wing curvature, patterns, chair leg shapes, and overall style; the defendants also conducted designs adapted for industrial production.

Table 9: Rubrics of Task 2: Practical Case Analysis (I).

Dimension	Content	Point Allocation	Examples
Reasoning Process	Non-Sequential Reasoning	Points are awarded for individual analytical points regardless of their order (10 Points per Step/Point).	[Reasoning Process Scoring] Analyze the four requirements according to Article 122 of the Civil Procedure Law: Analysis regarding Plaintiff's Standing (+10 points) • Mention that Party A is the owner of the involved Xiaomi mobile phone. • Mention that Party B possesses the phone and refuses to return it, directly infringing upon Party A's ownership. Analysis regarding the Specificity of the Defendant (+10 points) • Mention that Party A's target of litigation is Party B, an individual industrial and commercial household providing screen replacement services. • Mention that Party B's identity is definite, and the contact information and business address are clear (Party A was able to mail the phone to their repair station). Analysis regarding Specific Litigation Requests, Facts, and Grounds (+10 points) • Conduct an analysis of the specific litigation requests: Mention the litigation request"to order Party B to immediately return the involved Xiaomi mobile phone and bear the litigation costs of this case." • Conduct an analysis of the specific facts and grounds: Centering on the litigation requests, Party A stated a series of specific facts, including the establishment of a repair contract relationship between the two parties, the accidental mailing of the phone by themselves, and the detention of the phone by Party B without return. Party A claims that Party B's behavior is"without legal basis" and constitutes the legal grounds for the lawsuit. Simultaneously, Party A submitted preliminary evidence such as chat records and courier vouchers to support the claim. Analysis regarding the Scope of Court Acceptance and Jurisdiction (+10 points) • Mention that this case is a dispute arising from the possession and return of property, which is a property rights dispute and a typical civil dispute. • Mention that according to the prompt information,"Party A filed a lawsuit in a court with jurisdiction," which indicates that the issue of jurisdiction has been resolved and requires no further in-depth discussion.
Statute Location	Evaluation of Statute Location	Accurately citing the relevant law or specific article earns +5 points.	(+5 points) Article 23 of the Company Law of the PRC: Where a shareholder of a company abuses the independent status of the company as a legal person and the limited liability of shareholders to evade debts and seriously damages the interests of the creditors of the company, the shareholder shall bear joint and several liability for the debts of the company. Where a shareholder uses two or more companies under their control to commit the acts specified in the preceding paragraph, each company shall bear joint and several liability for the debts of any of the companies. In the case of a company with only one shareholder, if the shareholder cannot prove that the company's property is independent of the shareholder's own property, the shareholder shall bear joint and several liability for the debts of the company.

Table 10: Rubrics of Task 2: Practical Case Analysis (II).

C.3 Rubrics of Legal Document Generation Task

In rubric design, we reduce the weight assigned to formal presentation to less than 10% and focus evaluation on substantive content. We emphasize models' ability to filter incorrect and irrelevant information and design case-specific rubrics for each scenario to ensure objective and effective evaluation of model outputs.

Specifically, the rubric of Plaintiff focuses on:

- Procedures And Jurisdiction: Identify the Applicable Court, Determine Proper Party, Select the most Favorable Cause of Action
- Key Claims and Best Strategy of Claim: Ensuring Reasonableness and Completeness of the Prayer for Relief, Capability to Accurately Calculate Amounts
- The Construction of Facts by Evidence: Logical Structuring of Facts and Reasoning, Preliminary Organization of Evidence

The rubric of the Defendant focuses on:

- Selection of Defense Strategies: Restructuring of Legal Relationships, Determining Core Defense Strategies, Demonstrates Solid and Coherent Argumentation
- Procedural Defenses of Jurisdiction and Admissibility: Jurisdictional Objections, Challenges to Legal Standing, Defenses Based on Procedural Timelines, Procedural Dismissal Motions
- Defenses Against Damages and Compensation Claims: Arguing of causation or intervening factors, Excluding non-compensable losses such as indirect damages, Challenging the calculation of plaintiff's claimed amount

In addition, to ensure compliance with the fundamental normative requirements of legal documents, a set of uniform penalty criteria has been designed and shall apply to all types of assessed documents, such as statements of claim and defense. The uniform penalty criteria follow a cumulative deduction rule, with deductions accumulating up to the total score of the question. The specific criteria are as follows:

- Accuracy of Provisions Citation: The serial numbers and content of legal provisions cited in the document must be authentic, valid, and accurate. Each error in a legal provision (e.g., incorrect serial number, misrepresentation of content, citation of an invalid provision, etc.) will result in a 10-point deduction.
- Compliance with Instructions: The document must be generated strictly in accordance with user instructions or the template of the specified document type. If the document type generated does not match the required type (e.g., a statement of defense is requested but a statement of claim is produced), a 10-point deduction will apply.
- Document Standards: The following standards must be met simultaneously. For each standard not satisfied, a 10-point deduction will apply:
  - The format complies with statutory format requirements for the corresponding type of legal document.
  - Wording is precise and standardized, using formal legal terminology without idiom or ambiguous expressions.
  - All essential elements of the document are included, with no omission of core components (e.g., absence of party information, or lack of a clear statement of defense).

Dimension	Content	ID/Point
The Rubric of Plaintiff		
Procedures And Jurisdiction	• Identify the Applicable Court	Following with the Civil Procedure Law and relevant rules, accurately determine the court with territorial and hierarchical jurisdiction to ensure the litigation proceedings.
	• Determine Proper Party	Clearly identify the plaintiff's and defendant's legal interest, ensure the legitimacy of the parties' standing, and guarantee the accuracy and completeness of their information.
	• Select the most Favorable Cause of Action	Precisely define the legal relationships, select suitable cause of action, and prioritize the cause of action most favorable to one's own side.
Key Claims and Best Strategy of Claim	• Ensuring Reasonableness and Completeness of the Prayer for Relief	Translate the client's interests into specific and enforceable claims, comprehensively covering core rights and ancillary rights.
	• Capability to Accurately Calculate Amounts	Accurately calculate the amounts related to the claims, ensuring that the calculation basis, standards, and commencement and cutoff dates are lawful, and that the process is clear and verifiable.
The Construction of Facts by Evidence	• Logical Structuring of Facts and Reasoning	Narrate the case facts clearly along a timeline, highlighting key events, and achieve precise alignment between facts and legal grounds.
	• Preliminary Organization of Evidence	Organize the evidence list, clarify the purpose of proof, and ensure that each piece of evidence corresponds to the factual statements, without omitting any core or key evidence.
The Rubric of Defendant		
Selection of Defense Strategies	• Restructuring of Legal Relationships	Break through the legal relationship framework established by the plaintiff's claims, reconstruct a legal relationship favorable to one's own side.
	• Determine Core Defense Strategies	Identify key winning arguments in the case. For example: arguing for the invalidity of the contract, claiming force majeure or fundamental change of circumstances, or invoking the defense of prior performance.
	• Demonstrates Solid and Coherent Argumentation	Ensure that all defense arguments are consistent and free of contradictions, and that a complete logical chain from "facts → law → conclusion" is formed, thereby guaranteeing the coherent argumentation.

Table 11: Rubrics of Task 3: Legal Document Generation (I).

Dimension	Content	ID/Point
Procedural and Jurisdictional Defenses	• Jurisdictional Objections	Identify jurisdiction errors or exclude court jurisdiction based on arbitration agreements; accurately distinguish the jurisdictional boundaries of criminal-involved cases to secure procedural advantages.
	• Challenges to Legal Standing	Examine and challenge the legitimacy of the plaintiff's standing to sue and the propriety of the defendant as a party.
	• Defenses Based on Procedural Timelines	Identify defense opportunities based on key time points such as the statute of limitations and exclusion periods.
	• Procedural Dismissal Motions	For example, apply for a suspension of proceedings (such as in cases where criminal proceedings take precedence over civil ones, or while awaiting the outcome of another case), or argue that the plaintiff has not exhausted prerequisite remedies (such as the confirmation procedure in state compensation claims).
Defenses Against Damages and Compensation Claims	• Arguing of causation or intervening factors	Through evidence or argumentation, sever the direct causal link between one's own actions and the alleged harm, or assert that intervening factors caused the harm.
	• Excluding non-compensable losses such as indirect damages	In accordance with legal provisions, exclude non-compensable losses such as indirect losses and lost profits, and confine the scope of compensation to reasonable limits.
	• Challenging the calculation of plaintiff's claimed amount	Challenge the calculation basis, standards, or other elements of the amount claimed by the plaintiff, assert applicable rules such as set-off of gains and losses, and seek to reduce the compensation amount.
The Rubric of Penalty		
Penalty Item	**Specific Requirements**	
Accuracy	Accurately citing the specific article	The serial numbers and content of legal provisions cited in the document must be authentic, valid, and accurate. Each error in a legal provision (e.g., incorrect serial number, misrepresentation of content, citation of an invalid provision, etc.) will result in a 10-point deduction.
Compliance with Instructions	Following formatting and submission rules	The document must be generated strictly in accordance with user instructions or the template of the specified document type. If the document type generated does not match the required type (e.g., a statement of defense is requested but a statement of claim is produced), a 10-point deduction will apply.
Document Standards	• Professional tone and language	The format complies with statutory format requirements for the corresponding type of legal document.
	• Logical document structure	Wording is precise and standardized, using formal legal terminology without idiom or ambiguous expressions.
	• Proper legal formatting	All essential elements of the document are included, with no omission of core components (e.g., absence of party information, or lack of a clear statement of defense).

Table 12: Rubrics of Task 3: Legal Document Generation (II).

D Model Details

In this section, we provide brief introductions to the 10 large language models evaluated in our experiments.

Claude-Sonnet-4.5 (Anthropic, 2025) The latest iteration in Anthropic's Claude series, Sonnet-4.5 balances high performance with cost-efficiency, demonstrating enhanced capabilities in nuanced instruction following and complex reasoning tasks.

Gemini-2.5-Pro (Comanici et al., 2025) An upgraded version of Google's multimodal model, Gemini-2.5-Pro features improved long-context understanding and more robust performance across diverse logic and coding benchmarks.

Gemini-3.0-Pro-Preview (Pichai et al., 2025) The next-generation preview model from the Gemini family, designed with advanced native multimodal integration and superior reasoning depth for complex agentic tasks.

GPT-4o-20240806 (Katz et al., 2024) A highly optimized snapshot of OpenAI's GPT-4o, known for its rapid inference speed and state-of-the-art performance in multimodal processing and structured output generation.

GPT-5-0807-Global The foundational version of the GPT-5 series, representing a significant leap in general intelligence, with substantial improvements in world knowledge and reasoning stability compared to its predecessors.

GPT-5.2-1211-Global An iterated and refined version of the GPT-5 architecture, offering enhanced instruction adherence and reduced hallucination rates, specifically tuned for complex problem-solving.

DeepSeek-V3.2 (Liu et al., 2025a) Developed by DeepSeek, this model utilizes a Mixture-of-Experts (MoE) architecture to achieve competitive performance with top-tier proprietary models while maintaining high inference efficiency.

Kimi-K2 (Team et al., 2025) The second-generation model from Moonshot AI, Kimi-K2 specializes in lossless long-context processing, capable of handling massive information retrieval and synthesis tasks effectively.

Qwen3-235b-Think (Yang et al., 2025) The latest version of the Qwen model family. We use qwen3-235b-a22b-think, a flagship model featuring a unified framework that dynamically allocates a "thinking budget" to balance latency and performance, specifically optimized for complex multi-step reasoning.

Qwen3-Max The most capable model in Alibaba's Qwen3 lineup, designed to handle the most demanding tasks with comprehensive knowledge coverage and superior complex agent planning abilities.

E Prompt Engineering

We utilized specific prompts to guide the models in generating responses and to instruct the Judge model for evaluation. Tables 13, 15, 16,through 17, 18, 19 display the core content of these prompts.

E.1 Response Generation Prompts

E.1.1 Prompt Design for Task 1: Public Legal Consultation

The prompt for Task 1 transitions the LLM from a passive information provider to an active legal interrogator by enforcing a rigorous fact-finding protocol (As shown in Tables 13). This design requires the model to reconstruct the full scope of an event by prioritizing the Seven Elements of Evidence, which include time, location, parties, actions, amounts, evidence, and sources. To address the inherent bias in user narratives, the prompt incorporates a Concealment Detection mechanism that utilizes adversarial and reciprocal questioning to probe for latent contradictions or unfavorable facts. Finally, the model is mandated to maintain a logical hierarchy by addressing high-stakes issues such as personal safety and property title before supplementary details. This structured approach ensures professional neutrality and provides the granularity necessary for complex litigation strategy.

This is an example of adapted case (Shown in Table 14). The adapted case incorporates subtle factual distortions and selective disclosures to obfuscate the legal nature of the relationship between the parties. By introducing personal affiliations (the romantic relationship), emphasizing formalistic but non-determinative evidence (access to the corporate seal and third-party emails), and detailing a circular financial arrangement for social insurance payments, the narrative creates a facade of a legitimate labor relationship. These "traps" are designed to test the analyst's ability to look beyond surface-level administrative records and identify the absence of substantive labor elements, such as organizational subordination, actual wage payment by the employer, and the existence of a bona fide job position.

E.1.2 Prompt Design for Task 2: Practical Case Analysis

The prompt for Task 2 evaluates the ability of Large Language Models to perform systematic legal deconstruction by enforcing a structured cognitive framework aligned with professional judicial standards (As shown in Tables 15). This design mandates a strict "Conclusion First" response order, followed by factual extraction, logical reasoning, and statutory citations to mirror the organizational style of professional legal memorandums. This structured reasoning chain creates multiple points of verification, ensuring granular transparency and allowing practitioners to identify specific points where a model might diverge from established judicial logic. Furthermore, the step-by-step requirement necessitates a direct alignment between case facts and specific legal provisions, providing an explicit mapping that allows experts to audit the precision of the reasoning. This mechanism effectively addresses the reasoning gap by preventing models from reaching correct conclusions through flawed or hallucinated legal paths.

E.1.3 Prompt Design for Task 3: Legal Document Generation

The prompt for Task 3 evaluates the proficiency of Large Language Models in professional legal drafting by requiring the transformation of informal client narratives into formal legal instruments (As shown in Tables 16). A primary feature of this design is the requirement for autonomous document classification, where the model must independently determine the appropriate document type, such as a Bill of Complaint or a Statement of Defense, based on the procedural roles identified within the context. This requires the model to distinguish between the grievances of a plaintiff and the rebuttals of a defendant before applying the corresponding legal template. Furthermore, the prompt mandates a critical analysis of user

input to identify and correct factual inconsistencies or potential traps resulting from low legal literacy. By filtering out irrelevant emotional descriptions while retaining legally significant facts, the model ensures that the final document remains legally sound and compliant with the statutory standards and judicial norms of Mainland China.

Task 1 Prompt: Public Legal Consultation

ROLE

You are a senior attorney with over a decade of practice experience, possessing expert-level knowledge of current Chinese laws, regulations, and judicial practices. You excel at reconstructing the full scope of an event based on the one-sided facts provided by a client, identifying potential concealments or biases, and completing the evidentiary chain and timeline to prepare for subsequent legal analysis and litigation strategy.

CORE REQUIREMENTS

1. **Neutrality:** Remain objective and non-suggestive. Avoid taking sides or using qualitative/judgmental labels (e.g., "scumbag," "abuser," or "certainly").

2. **Verifiability and Evidence-Orientation:** Focus inquiries on "Time, Location, Parties, Actions, Amounts, Evidence, and Sources."

3. **Comprehensive Coverage:** Cover all aspects of the event, details of signed agreements, mutual faults, and potential risk points.

4. **Prioritize Critical Issues:** Address questions that most significantly impact the case outcome first (e.g., safety, evidence, property ownership, nature of debt, or custody arrangements) before moving to supplementary details.

5. **Multi-Question Output:** Provide at least 25 questions per response in a numbered list. Questions must be concise, clear, and focused on a single point of inquiry.

6. **Identify Contradictions or Gaps:** You must include clarification questions to address latent contradictions (e.g., inconsistent timelines, unclear sources of funds, or questionable authenticity of agreements).

7. **Concealment Detection:** Proactively inquire about potentially omitted unfavorable facts using "reciprocal questioning" (e.g., "Was it a mutual physical altercation?", "Who initiated the physical contact?", "Are there police records?", "Was there an extramarital affair?", "Was the loan jointly signed?").

8. **Output Format:** Provide only the list of questions. Do not include any introductory or concluding text.

RESPONSE EXAMPLE

1. Does [Name] hold a Special Operation Certificate for welding?

2. In the agreement regarding "[Subject]" between [Party A] and [Party B], was the relationship defined as employment or a dynamic contract for work?

3. After the 2,000 RMB was transferred to [Name]'s personal WeChat account, was there a written agreement stating this was "collection on behalf of the company" or any subsequent vouchers for company receipts?

4. Does the outer packaging of the tea involved in this case label the entrusting party as [Company X] and the entrusted party as [Factory Y]?

5. .....

6. .....

7. .....

USER INPUT FORMAT

"You are an attorney. Your client is consulting with you and states: `<conversation>`. You are now required to raise follow-up questions to your client to determine the full scope of the event."

Table 13: System prompt for Legal Consultation task.

Table 14: Comparison of Case Details

Raw User's Narrative Version	Bar Exam Version
It all started back in November 2017. At that time, I was introduced to Hengyuan International Trade Co., Ltd. by a friend named Lin Tao, who had grown up with me since childhood (Hidden case detail: Lin Tao and the party involved were in a romantic relationship; this detail has no impact on the case determination, but adding this information makes it easier to identify the fabricated labor relationship). He and I were extremely close—our parents were old neighbors—and he was already a shareholder in the company at that time, holding 30% of the shares, so I trusted him completely. He told me that the company had just started and urgently needed a reliable person to help the general manager handle core affairs, so he had me take the position of general manager's assistant. On my first day, the general manager personally received me and talked with me in his office for nearly an hour. He explicitly said: 'Your salary will be 16,000 yuan per month. We'll settle on it verbally for now, and the contract will be supplemented next week.' (No such position actually existed; this was a statement colluded between Lin and the manager. Adding this case detail makes it easier to identify the fabricated labor relationship.) I also saw the company's business license and official seal on his desk, which made me feel that everything was very formal and legitimate. From that day on, I devoted myself wholeheartedly to the work, going out early and returning late every day, helping the general manager organize customer data, draft meeting minutes, and coordinate with three key foreign trade customers—these customers even sent me thank-you emails later! (The so-called 'coordinating with customers' was actually business from another company controlled personally by Lin Tao, unrelated to Hengyuan Trade, but the case description deliberately blurred this as 'company customers.' Adding this case detail makes it easier to identify the fabricated labor relationship!)	In November 2017, Qiao was introduced by her boyfriend Lin Tao (who was also a shareholder of Hengyuan International Trade Co., Ltd. [hereinafter "Hengyuan Company"], holding 30% of shares) to join Hengyuan Company as a general manager's assistant. During her employment, Qiao was responsible for organizing customer data, drafting meeting minutes, and coordinating with three foreign trade customers (which were actually business operations controlled personally by Lin Tao). However, Hengyuan Company did not implement attendance tracking for Qiao, did not issue her an employee badge or system account, and did not sign a labor contract with her.

Continued on next page...

4

Raw User's Narrative Version	Bar Exam Version
But who could have imagined that not a single penny of salary was ever paid? At first, Lin Tao comforted me, saying: 'The company's cash flow is tight, but I will definitely get your social insurance set up first—don't worry.' (The social insurance fees were actually transferred in cash from Qiao Xiaoli to Lin Tao, who then paid them on behalf of the company in the company's name; the company did not actually bear the costs. Her pregnancy at the time of employment should be noted for disclosure.) Sure enough, from November 2017 to June 2018, my social insurance was normally paid. I checked the system records myself—the insured unit clearly stated 'Hengyuan International Trade Co., Ltd.' Doesn't this prove that I was their employee? In early 2018, I found out I was pregnant and immediately told the general manager. He even congratulated me, saying, 'The company supports you—take your maternity leave according to regulations, and your salary will not be reduced by a single cent.' I started my maternity leave at the end of April and gave birth to my baby smoothly in May (I was already pregnant at the time of employment, in order to qualify for maternity subsidies and protection). But up to now, not a single penny of maternity leave salary has been received. When I asked the finance department, they stammered and said, 'We haven't received any salary payment instructions.' (No maternity certificates, maternity leave application forms, or other required documents were submitted to the company.) [Omission due to length]	From November 2017 to June 2018, at Qiao's request and with her bearing the social insurance costs herself, Hengyuan Company normally paid social insurance for Qiao, with the social insurance system showing Hengyuan Company as the participating unit. In early 2018, Qiao informed the general manager about her pregnancy, and she began maternity leave at the end of April. After giving birth in May, Hengyuan Company did not pay her maternity leave wages. On July 12, 2018, the general manager notified Qiao via WeChat to terminate the labor relationship on the grounds of "optimizing personnel structure," without providing a written termination notice. Thereafter, Hengyuan Company stopped paying Qiao's social insurance and refused to process social insurance reduction procedures, causing Qiao to miss two job opportunities. Qiao filed claims requesting the company to pay wages, double salary difference for unsigned contract, and other amounts totaling 290,000 yuan, and to issue a certificate of termination.
Question: Based on the above case details, analyze whether each of Qiao's litigation claims can be supported, and explain the reasons.	

Table 15: System prompt for Case Analysis task.

Task 2 Prompt: Practical Case Analysis

ROLE

You are a legal practice expert with over a decade of experience, possessing mastery over current Chinese laws, regulations, and judicial practices. You specialize in deconstructing complex legal issues into clear logical modules and providing answers strictly adhering to a professional style characterized by "Conclusion First, Facts as the Foundation, Rigorous Reasoning, and Statutory Support."

CORE REQUIREMENTS

1. **Strict Sequence:** The response must unfold in the following order using the corresponding headings: `Conclusion`, `Case Facts`, `Reasoning Process`, `Statutory Basis`.

2. **Content Standards:**

• **Conclusion:** Provide a direct and clear affirmative or negative judgment regarding the core dispute of the inquiry.

• **Case Facts:** Based on the user's account, extract the factual elements relevant to the legal judgment concisely and objectively. Do not add assumptions or conjectures.

• **Reasoning Process:** Demonstrate step-by-step how the facts link to the statutes to derive the conclusion. Use a layered or bulleted format.

• **Statutory Basis:** Cite current and effective laws, judicial interpretations, or administrative regulations of Mainland China. Specify the issuing authority, title, and specific article, paragraph, or item. Provide the original text of the statute.

TEMPLATE

`Conclusion`

[Provide a clear and definite legal judgment here]

`Case Facts`

(Extract only key facts relevant to the conclusion; remain objective and concise)

`Reasoning Process`

1. .....;

2. .....;

`Statutory Basis`

Article X, Paragraph X of the XXX Law (Issuing Authority: XXX)

(If multiple bases exist, arrange them by hierarchy of authority or relevance)

RESPONSE EXAMPLE

`Conclusion`

1. The matrimonial property relationship shall be governed by the laws of the Macau SAR...

2. Property I (Mainland City A) is identified as joint matrimonial property.

3. Property II (Mainland City B) is identified as the personal property of Party R...

`Case Facts`

• **Key Fact 1:** Party L (Macau resident) and Party R (Mainland resident) registered marriage in Macau (2013) and divorced (2016)...

• **Key Fact 2:** Signed "Prenuptial Agreement Record" stipulating Macau as habitual residence and "General Community of Property"...

• **Key Fact 3:** Property I purchased by Party L before marriage, registered solely to Party L...

• **Key Fact 4:** Property II purchased by Party R's parents via housing reform; Party L consented to registration under Party R...

`Reasoning Process`

Step 1: Determine applicable law. According to Art. 24 of the *Law of the Application of Law for Foreign-related Civil Relations*: parties may choose the law of habitual residence. Parties chose Macau; thus, Macau law applies.

Continued on next page...

Table 15 – Continued from previous page

Task 2 Prompt: Practical Case Analysis (Continued)

Step 2: Determine property regime. Macau Civil Code Art. 1609: community property consists of all present/future property unless excluded by law.

Step 3: Ownership of Property I. Under "General Community," pre-marital purchase (Property I) is not excluded; thus, it is joint property.

Step 4: Ownership of Property II. Involves seniority welfare. Party L's waiver indicates recognition of title. Per Macau Civil Code Art. 1610(1)(a), property gifted with exclusion clauses is personal. Property II is Party R's personal property.

`Statutory Basis`

- Article 24 of the *Law of the Application of Law for Foreign-related Civil Relations of the PRC*...
- Article 1609 of the *Civil Code of Macau*...
- Article 1610 of the *Civil Code of Macau*...

USER INPUT FORMAT

<questions> Please provide your answer in the following format: `Conclusion+Case Facts+Reasoning Process+Statutory Basis` <asking>

Task 3 Prompt: Legal Document Generation

ROLE

You are a legal practice expert familiar with the formal drafting standards for legal instruments. Based on the statements or information provided by your client (party involved), you are required to generate a legal document.

1. DOCUMENT TYPES

Please generate the appropriate legal document based on the following options (ensure the content meets the specific requirements of the selected type):

(1) Application (Bill of Complaint)

(2) Defendant (Statement of Defense)

2. DOCUMENT REQUIREMENTS

1. **Template Adherence:** Drafting must automatically follow the specific template provided for each case. If no specific template exists for a particular cause of action, use the general standard legal format used in Mainland China.

2. **Length Constraint:** The generated legal document must be between 2,500 and 3,000 characters in length.

3. **Concise Output:** Output the document text only. Do not include any "fluff," such as conversational filler or simulated lawyer-client dialogue (e.g., "Certainly, here is your document...").

4. **Autonomous Selection:** You must independently determine which of the three document types is appropriate based on the provided prompt and context.

5. **Critical Analysis:** The client may have low legal literacy, and their statements may not be entirely accurate or may contain errors. You must identify potential "traps" or factual inconsistencies and draft a document that is legally sound, reasonable, and compliant.

6. **Statutory Citations:** You must cite authentic and currently effective laws and regulations in the document.

USER INPUT FORMAT

"You are an attorney. Your client has approached you and described the situation as follows: `<conversation>`. Based on the current scenario, please draft a Complaint / Answer / Judicial Decision for your client."

Table 16: System prompt for Document Generation task.

E.2 Judge Model Scoring Prompts

E.2.1 Scoring Prompt 1: Public Legal Consultation

To provide an objective and scalable assessment of model performance, we developed a standardized scoring protocol for Task 1 that transitions the evaluator into a neutral judge (As shown is 17). This process is grounded in a strict literal comparison between the generated questions and expert-curated rubrics, ensuring that the final score directly reflects the presence of specific legal inquiry points rather than subjective interpretations. The mechanism employs a nuanced matching logic where full credit is awarded for complete semantic consistency, while partial credit is granted for responses that capture core themes but lack specific qualifiers. Furthermore, the protocol enforces strict neutrality and mandates a structured analysis for every scoring item.

E.2.2 Scoring Prompt 2: Practical Case Analysis

The evaluation protocol for Task 2 employs a rigorous item-by-item inspection mechanism to assess the quality of complex legal case analyses (As shown is 18). This system requires the evaluator to segment model responses into four distinct functional components comprising the conclusion, statutory basis, case facts, and reasoning process. By applying specialized rubrics to each section independently, the judge adheres to a strict literal matching principle that awards points only for explicitly mentioned information while prohibiting subjective inferences or supplementary reasoning on behalf of the model. The scoring workflow follows a structured deconstruction of the legal syllogism, ensuring that the major premise, minor premise, and logical application are each validated against expert criteria. This granular approach is finalized through a standardized JSON output format that details the breakdown of points, specific textual mentions, and analytical rationales.

E.2.3 Scoring Prompt 3: Legal Document Generation

The scoring protocol for Task 3 evaluates the proficiency of Large Language Models in generating formal legal instruments that adhere to professional drafting standards (As shown is 19). This evaluation process assesses the model's ability to autonomously select the correct document type, such as a Bill of Complaint, Statement of Defense, or Judicial Decision, based on the procedural context of the user narrative. The judge measures the output against stringent formal requirements, including strict adherence to Mainland China legal templates, authentic statutory citations, and a specific length constraint of 2,500 to 3,000 characters. Furthermore, the scoring mechanism accounts for the model's capacity for critical analysis, specifically its ability to identify and rectify factual inconsistencies or legal traps within informal client statements. By comparing the generated document against expert-curated rubrics, the evaluator ensures that the final output is not only concise and devoid of conversational filler but also legally sound and compliant with judicial norms.

Scoring Prompt 1: Public Legal Consultation

ROLE

You are an evaluator. Your task is to score a model's output by comparing it against a standard answer containing specific scoring items to assess the model's performance.

INPUT AND OUTPUT FORMATS

Input consists of two parts:

1. Agent's Question List: The list to be evaluated (10–25 numbered items).

2. Standard Rubric: (Total Score: N points; each point is formatted as "(+X points) Point Description").

Output must follow this fixed format:

The response score is XX points. Total score is XX, with a scoring rate of XX%.

The analysis is as follows:

• Point 1: [Scoring Status], Rationale: .....

• Point 2: [Scoring Status], Rationale: .....

• .....

SCORING RULES

1. **Literal Comparison:** You do not need to judge the accuracy of the answer itself. Your sole task is to strictly follow the standard answer, determine whether the model's response contains identical or substantively identical content, and calculate the final score.

2. **Strict Neutrality:** Do not suggest new questions, provide legal advice, evaluate the parties involved, or restate lengthy background information. Provide only the comparison and scoring explanations.

3. **Matching Logic:** For each "Question Point" in the standard rubric, search for the best match within the agent's question list. Award points based on the degree of matching:

o 100% Score: If the response fully covers key elements and is semantically consistent.

o Partial Credit: If the response hits the core theme but lacks key qualifiers. For example, if the standard requires "police records, hospitalization records, and medical certificates" but the response only asks about "police records," award partial credit at your discretion.

o 0% Score: If the response fails to cover the semantic key points.

Summary: Aggregate the scores and calculate the Scoring Rate = Earned Points / Total Points.

USER INPUT

The consultation scenario between the attorney and the client is `<conversation>`. In this scenario, the scoring rubric for the attorney's performance is `<rubrics>`. Please evaluate the following questions posed by the attorney: `<answers>`.

Table 17: Judge prompt for scoring Task 1: Public Legal Consultation.

Scoring Prompt 2: Practical Case Analysis

ROLE

You are a rigorous legal practice evaluation expert, specializing in item-by-item inspection of legal practice responses based on precise grading criteria. You strictly adhere to the principle of "Award points if mentioned; deduct points if omitted," and you must avoid making subjective inferences.

CORE TASK

Evaluate legal practice responses accurately according to the provided detailed scoring rubrics. A response is divided into four sections: Conclusion, Statutory Basis, Factual Summary, and Reasoning Process. Each section may have its own independent rubric.

INPUT INFORMATION STRUCTURE

Question Content: A brief case description and the specific inquiry.

Scoring Rubric: Divided into multiple parts. The format for each may follow this example:

- Conclusion Rubric
- Statutory Basis Rubric
- Case Facts Rubric
- Reasoning Process Rubric
- Other Section Rubrics

Response to be Evaluated: The text requiring scoring, clearly labeled with the four sections: "Conclusion," "Statutory Basis," "Case Facts," and "Reasoning Process."

SCORING PRINCIPLES

1. **Strict Literal Matching:** Use the phrasing in the rubric as the standard; check whether the response contains identical or substantially identical expressions.
2. **Independent Evaluation:** Each scoring item is calculated independently without considering other items.
3. **No Inference or Supplementation:** Score based solely on what is explicitly mentioned in the response; do not perform reasoning or add information on the model's behalf.
4. **Clear Addition/Deduction:** Strictly follow the assigned values for adding or deducting points.
5. **Discretionary Partial Credit:** If the response does not fully cover a scoring point, award partial credit based on the key elements that were successfully addressed.

SCORING WORKFLOW

Step 1: Segment the Response. Divide the response into four parts: Conclusion, Statutory Basis (Major Premise), Case Facts (Minor Premise), Reasoning Process.

Step 2: Individual Section Scoring. For each part, use the corresponding rubric. The steps for each section are: (1) Analyze Rubric: Deconstruct the rubric into scoring items and deduction rules. (2) Content Inspection: Check if each required item from the rubric is mentioned in the response.

Step 3: Structural Scoring (If applicable). If the rubric includes criteria for overall structure, evaluate the response accordingly.

Step 4: Summary and Feedback.

OUTPUT FORMAT

Output a detailed evaluation for the following four parts, along with the total score and overall feedback.

```
{
  "score_details": {
    "Statutory Basis": {
      "total_points": 0, "max_points": 0,
      "breakdown": [ {
        "rubric_item": "Description...", "max_points": 0, "points_awarded": 0,
        "mentions": ["Relevant snippets..."], "rationale": "Reasoning..."
      } ]
    }, ... [Structure repeated for Case Facts, Reasoning Process, Conclusion, Others]
  },
  "total_score": { "total_awarded": 0, "total_max": 0, "percentage": "0%" },
  "overall_feedback": { "strengths": [], "weaknesses": [], "suggestions": [] }
}
```

USER INPUT

The case analysis question is `<question><asking>` and its scoring rubric is `<rubrics>`. Please evaluate the following response: `<answers>`.

Table 18: Judge prompt for scoring Task 2: Practical Case Analysis.

Scoring Prompt 3: Legal Document Generation
ROLE You are a legal practice expert specializing in the formal drafting standards of legal instruments. I will provide you with statements or other information from your client (the party involved) (${question}). Based on this information, you are required to generate a formal legal document. The document types and specific requirements are as follows:
1. DOCUMENT TYPES Generate the appropriate legal document from the following options (ensure the content strictly adheres to the specific requirements of that type): (1) Application (Bill of Complaint) (2) Defendant (Statement of Defense) (3) Judicial Decision (Judgment/Ruling)
2. DOCUMENT REQUIREMENTS (1) **Template Adherence:** Automatically draft the document according to the specific template provided for the given case. If no specific template for the cause of action is available, use the standard universal legal format used in Mainland China. (2) **Length Constraint:** The legal document you draft must be between 2,500 and 3,000 characters. (3) **Concise Output:** Output the document text only. Do not include any "fluff," such as conversational filler or simulated lawyer-client dialogue. (4) **Autonomous Selection:** You must independently determine which of the three document types is appropriate based on the provided prompt and context. (5) **Critical Analysis:** Clients may have low legal literacy, and their statements may be inaccurate or contain biases. You must identify potential "traps" and factual inconsistencies to ensure the final document is legally sound, reasonable, and compliant. (6) **Statutory Citations:** You must cite authentic and currently effective laws and regulations in the document.
USER INPUT This is a scenario for legal document drafting: <conversation>. The evaluation criteria are <rubrics>. Please score the following legal document: <answers>.

Table 19: Judge prompt for scoring Task 3: Legal Document Generation.

F Evaluation Methodology Details

F.1 Judge Model Validation Results

To identify the most accurate judge model, we calculated the alignment between scores from human experts and various LLMs to determine which model most closely approximates human judgment. The evaluation dataset consisted of the following:

1. **Case Analysis:** A random 10% sample of all items, paired with responses from a randomly selected model.

2. **Document Drafting:** Ten cases each for Complaints and Answers, paired with responses from a randomly selected model.

3. **Legal Consultation:** Responses to 20 test items generated by four specific models (Claude 3.5 Sonnet, GPT-5-0807-Global, Qwen-Max, and Gemini 3 Pro Preview).

These responses were distributed to both human experts and LLMs for scoring. To ensure reliability, each response was evaluated by at least three human experts. We compared individual expert scores against the group average; the Pearson correlation coefficients between the three experts' scores and the human average ranged from 0.8391 to 0.8841, with a Mean Absolute Error (MAE) between 5.1937 and 7.9817. Table 4 presents the alignment results between candidate Judge models and human experts.

Following current research trends that employ high-performance models as evaluators, we tested Gemini 3 Pro Preview, GPT-5.1-1113-Global, and Qwen3-Max as potential judge models. Based on their comprehensive performance across different tasks, we selected **Gemini 3 Pro Preview** as our final judge model. Although it did not achieve the highest alignment in the Document Drafting task, it demonstrated superior performance in the remaining two categories.

F.2 Scoring Mechanics

To prevent scoring discrepancies, we utilize a scoring rate (Earned Points / Total Points). We established eight comprehensive metrics. Since LLMs are prone to calculation errors, we employ regular expressions to extract raw scores from the JSON or list outputs generated by the Judge model (as defined in the scoring prompts) and perform the final summation programmatically.

G Detailed Error Analysis

G.1 Error Analysis of Public Legal Consultation

For Task 1, we analyzed 183 low-scoring responses and identified four common categories of errors along with eight specific issues. The common problems observed in current LLMs during public legal consultation include:

- **Insufficient grasp of facts and causal relationships.** Models often struggle to accurately link reported events to their legal consequences.

- **Shallow understanding of evidence.** Models frequently fail to build a closed logical loop of questioning around evidence, missing the necessity of corroborating key facts.

- **Weak identification of specific evidence types.** There is a general lack of sensitivity to the specific forms of evidence required for different legal claims.

- **Defects in role adherence.** Some models exhibit limitations in their ability to discern and analyze the overall case context due to persona setting issues.

Beyond these commonalities, distinct model-specific weaknesses were observed. For instance, **GPT-5.2** tends to deviate from key facts during follow-up questioning, resulting in divergent lines of inquiry. **Gemini-3.0** struggles with cases involving complex evidence, often failing to understand probative requirements and asking questions without sufficient legal basis.

Overall, the performance of current LLMs in simulating real-world legal consultation remains insufficient. Whether in fact organization, causal inference, or evidence analysis, models generally exhibit shallow understanding and fragmented reasoning. This is particularly evident in complex cases with multiple causes or weak evidence chains, where misjudgments and omissions of key points are frequent.

G.2 Error Analysis of Practical Case Analysis

In Task 2, we examined 200 low-scoring cases and summarized the errors into four dimensions covering 16 specific issues. The common issues across models are:

- **Conclusion:** Fundamental errors in legal characterization.

- **Case Facts:** Omission of key facts that affect qualitative analysis, as well as loss of important details such as time, amount, and subject identity.

- **Reasoning Process:** Failure to follow the required legal reasoning sequence (e.g., syllogism).

- **Statutory Basis:** Absence of core legal provisions.

Model-specific analysis reveals that **GPT-5.2** has significant deficiencies in legal retrieval and application, frequently missing important legal bases or citing incorrect versions. **Gemini-3.0** shows defects in the integrity of reasoning chains and logical rigor, alongside serious issues with missing core statutes.

In summary, current LLMs have not yet reached a professional level in practical case analysis. While they can generate structured text, they exhibit obvious shortcomings in verifiable analysis: often omitting key facts, reasoning with improper order or broken chains, and misapplying laws (e.g., mixing new and old laws, or failing to distinguish between general and special provisions).

Case Study: Dispute over Cryptocurrency Mining Machine Sales
Case Summary: Party A purchased "Innosilicon AL0 Servers" (mining machines) from Party B for 50,000 RMB (48,000 RMB for goods + 2,000 RMB shipping). Party B labeled them "no returns or exchanges." Party A refused the initial delivery. A second notarized delivery confirmed the goods did not match the agreed model, and Party A refused again. Party B neither refunded nor replaced the goods. Both parties confirmed the subject matter was mining machines. Party B knew that "the state does not allow sales," and Party A was a "miner" aware of the usage.
Question: Answer with logic [Conclusion] + [Case Facts] + [Reasoning] + [Statutes]: Should the court order Party B to return 50,000 RMB and corresponding interest?
Analysis of Model Performance: The model exhibited significant failures in sequential reasoning:

- **Reasoning Order Violation:** The model skipped the mandatory first step of analyzing "contract validity" and directly proceeded to "breach of contract," assuming the contract was valid. It failed to follow the "Step 1 (Validity) → Step 2 (Return) → Step 3 (Interest)" sequence.
- **Missing Core Premise:** It failed to identify the illegality of trading mining machines, erroneously citing "no explicit legal prohibition" to deem the contract valid. This ignores regulatory restrictions and public order requirements.
- **Broken Logical Chain:**
  - In Step 2, instead of applying "void contract restitution" (Civil Code Art. 157), it applied "contract rescission for breach." It also failed to separate shipping costs from the goods price.
  - In Step 3, it supported the interest claim without analyzing the mutual fault regarding the illegal transaction, ignoring that both parties share responsibility for the void contract.

Diagnosis: The error stems from an inability to prioritize the "validity judgment" as a prerequisite. By skipping this foundational step, the subsequent reasoning on liability and remedies lost its legal basis.

G.3 Error Analysis of Legal Document Generation

For Task 3, an analysis of 183 low-scoring responses revealed two common issues:

- **Inaccurate Legal Application:** Insufficient grasp of the timeliness, relevance, and citation principles of laws.
- **Weak Jurisdiction Review:** Insufficient sensitivity and capability in reviewing jurisdictional rules.

These problems are prevalent across all models. Both international state-of-the-art models (e.g., GPT-5.2, Claude-4.5) and other strong models (e.g., DeepSeek, Qwen) struggle to perfectly resolve issues related to legal "hallucinations" and procedural logic.

Regarding individual differences, **GPT-5.2**, **Claude-4.5**, and **Gemini-3.0** make fewer errors in basic legal logic such as "proper party" determination and "fact summarization," indicating stronger instruction following and stability. Conversely, models like **DeepSeek**, **Kimi**, and **Qwen3-Max** are more prone to hallucinations and blindly following client instructions even when they are legally unsound.

Overall, while current models are mature in text organization, they lack substantive quality. They often fail to judge the suitability of subjects (blindly following wrong instructions), lack sensitivity to the statutory scope of claims (accepting unreasonable claims or omitting lawful ones), and violate the rigor required for legal documents.

Category	Error Type	GPT-5.2	Gemini-3.0	Claude-4.5	Qwen3-Max	Kimi-K2	DeepSeek-V3.2
Conclusion	Legal characterization error	✓	✓	✓	✓	✓	✓
	Incomplete conclusion (missing calcs)					✓	✓
	Validity/Time-effect error	✓	✓				
Facts	Omission of key facts	✓	✓	✓	✓	✓	✓
	Omission of details (time/amount)	✓	✓	✓		✓	✓
	Bias towards one party's claim		✓			✓	
Analysis	Incorrect reasoning order	✓	✓	✓	✓	✓	✓
	Incorrect analysis focus	✓		✓	✓	✓	✓
	Missing reasoning steps	✓	✓		✓	✓	✓
	Logical jumps/Non-rigorous	✓	✓		✓	✓	✓
Statutes	Missing core statutes	✓	✓	✓	✓	✓	✓
	Incorrect version (Old/Repealed)	✓	✓	✓	✓	✓	
	Correct citation, incorrect content	✓	✓		✓		
	Incorrect citation, correct content	✓					
	Improper application (General vs Special)	✓		✓	✓	✓	✓

Table 20: Distribution of error types in Task 2 (Practical Case Analysis) across evaluated models.

Category	Error Type	GPT-5.2	Gemini-3.0	Claude-4.5	Qwen3-Max	Kimi-K2	DeepSeek-V3.2
Legal Application	Inaccurate legal application (timeliness/relevance)	✓	✓	✓	✓	✓	✓
Procedural Review	Subject suitability error (blind compliance)				✓	✓	
	Insufficient sensitivity to jurisdiction rules	✓	✓	✓	✓	✓	✓
Substantive Review	Insufficient judgment of claim reasonableness				✓	✓	✓
	Lack of review on claimed amounts	✓	✓	✓	✓		✓
Drafting Standards	Unclear fact summarization				✓	✓	
	Failure to respond point-by-point (Defense)		✓	✓	✓	✓	✓

Table 21: Distribution of error types in Task 3 (Legal Document Generation) across evaluated models.