Pub-LawBench: Public-Oriented Benchmarking for LegalAI

Qiaoyu Zheng[1]*, **Zehan Ma**[2]*, **Yijing Zhang**[3]*, **Qiqi Wang**[1]*†, **Huijia Li**[1]†, **Qian Liu**[4]†
[1]School of Statistics and Data Science, AAIS, LPMC and KLMDASR, Nankai University, Tianjin, China
[2]School of Mathematical Sciences, AAIS, Nankai University, Tianjin, China
[3]School of Economics, AAIS, Nankai University, Tianjin, China
[4]School of Compute Science, The University of Auckland, New Zealand
{qiqi.wang, hjli}@nankai.edu.cn liu.qian@auckland.ac.nz
{2312302,mzhan22,yijingzhang}@mail.nankai.edu.cn

Abstract

Large language models (LLMs) are playing an increasingly pivotal role in LegalAI. However, existing benchmarks are primarily tailored for legal professionals, emphasizing deep reasoning and explainability. While public-facing legal applications demand outputs that are direct, actionable, and accessible, a need largely overlooked by current evaluation frameworks. To bridge this gap, we propose a public-oriented LegalAI benchmark grounded in legal functionalism and genre analysis. Specifically, we categorize public legal demands into two core tasks: *Instant Question Answering* and *Legal Text Generation*. We further introduce three public-oriented evaluation dimensions: *legal normativity*, *content relevance*, and *format usability*, which collectively assess the practical validity and user readiness of model outputs. To reflect real-world lay user usage, we evaluate 17 LLMs on Pub-LawBench using only simple prompts and Chain-of-Thought under a vanilla inference setting, excluding complex techniques like RAG or agent-based methods inaccessible to non-experts. Experiments reveal limitations of current LLMs in delivering effective public-oriented legal assistance, highlighting the need for more user-centric model development and benchmarking. [1]

1 Introduction

Recently, LLMs have significantly advanced LegalAI, enabling legal applications like charge prediction (Wu et al., 2023; Nigam et al., 2024), judgment summarization (Shukla et al., 2022), and contract review (Hendrycks et al., 2021). To support systematic evaluation, LegalAI benchmarks like LexGLUE (Chalkidis et al., 2022), LegalBench (Guha et al., 2023), LawBench (Fei et al., 2024), and CaseHOLD (Zheng et al., 2021) have been developed. However, they primarily target legal professionals (Kamigaito et al., 2023) and overlook lay users, thus hindering the development of publicly accessible LegalAI systems.

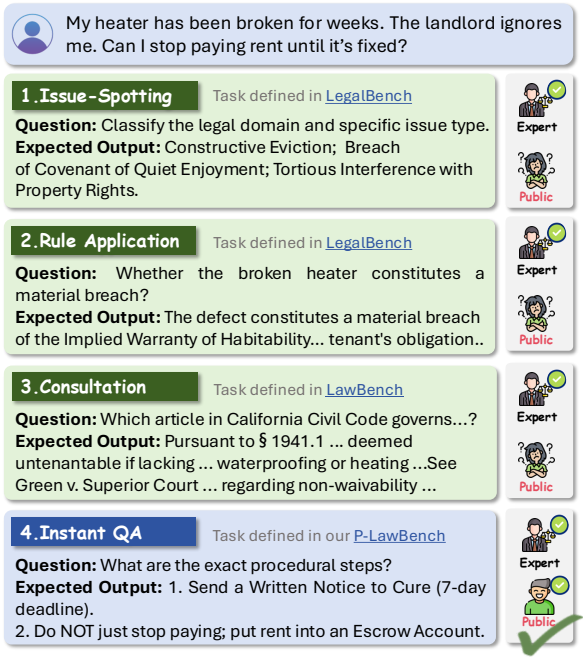

Figure 1: Comparison of profession- and public-oriented legal evaluation. Tasks 1–3, from LegalBench (Guha et al., 2023) and LawBench (Fei et al., 2024), evaluate legally precise, jargon-heavy outputs that satisfy professionals but confuse public users. Task 4 in our Pub-LawBench evaluates LLMs' ability to deliver clear, accessible output tailored for the general public.

Unlike professionals prioritizing rigorous reasoning, public users seek direct solutions like instant legal advice or functional document generation (Aryo Pradipta Gema et al., 2025). As shown in Fig. 1, *Issue-Spotting* is a professional-oriented task that assesses model identification of key legal issues (e.g., Constructive Eviction), a capability valued by experts but often opaque to lay users. Similarly, *Consultation* evaluates reasoning via citations and statutory references, yielding accurate

*Equal contribution.
†Corresponding author.
[1]Our code and datasets are available for review at https://github.com/Statistical-NLP-Lab/Pub-LawBench

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36281–36303
July 2-7, 2026

Tasks & Metrics		COLIEE	CaseHOLD	CUAD	LeCaRD	LexGLUE	LegalBench	LegalEval	LawBench	KoBLEX	GreekBar	SwissJudg	Pub-LawBench
Part I: Task Scenarios													
Instant QA	Legal Concept Interp.	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✓
	Elements of the Offense	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✓
	Statutory Interpretation	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓
	Open Q&A Solution	✓	✓	✗	✗	✓	✓	✗	✓	✓	✓	✗	✓
	Statute and Case Search	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓
	Classification Task	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗	✗	✓
	Prediction of the Verdict	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✓
Legal Gen	Fact Extraction & Summ.	✗	✗	✓	✗	✗	✓	✓	✓	✗	✗	✓	✓
	Reasoning Generation	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓
	Defense Generation	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
	Prediction of the Verdict	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓
Part II: Evaluation Dimensions													
Performance	Accuracy	✗	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✓
Content Relevance	Semantic Consistency	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Key Points Coverage	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓
	Info. Completeness	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✓
Legal Normativity	Statutory Citations	✓	✗	✗	✗	✓	✓	✗	✓	✓	✓	✗	✓
	Legal Judgments	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✗	✓
	Legal Language	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✓
Format	Structured Output	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓

Table 1: Comparison of **Pub-LawBench** with existing LegalAI benchmarks, including COLIEE (Rabelo et al., 2020), CaseHOLD (Zheng et al., 2021), CUAD (Hendrycks et al., 2021), LeCaRD (Ma et al., 2021), LexGLUE (Chalkidis et al., 2022), LegalBench (Guha et al., 2023), LegalEval (Kalamkar et al., 2023), LawBench (Fei et al., 2024), KoBLEX (Lee et al., 2025), GreekBarBench (Chlapanis et al., 2025), and Swiss-Judg (Rolshoven et al., 2025), in terms of task scenarios and evaluation dimensions. (✓: Supported; ✗: Not Supported)

yet impractical outputs. By contrast, our public-oriented *Instant QA* evaluates the models' delivery of direct, understandable guidance, such as "send a Written Notice to Cure", that empowers non-experts to take actionable steps.

In this work, we introduce Pub-LawBench, a public-oriented benchmark designed to evaluate the capacity of LLMs to support the practical legal needs of the general public. Specifically, we define two public-oriented task categories: *Instant Question Answering* (Instant QA) with 9 subtasks, aligning with legal functionalism and realism to prioritize actionable guidance; and *Legal Text Generation* (Legal Gen) with 4 subtasks, grounded in genre analysis and *Speech Act Theory* (Searle, 1969), viewing public legal texts as legally consequential acts requiring normative correctness and genre-appropriate structure.(Wang et al., 2025)

Moreover, current LegalAI metrics are largely performance-oriented, prioritizing legal precision but overlooking essential public needs. To truly serve non-experts, outputs must be understandable, actionable, and usable. We therefore propose public-oriented evaluation dimensions, i.e., *legal normativity*, *content relevance*, and *format usability*. Grounded in *Speech Act Theory* (Austin, 1962), these metrics mitigate the risks of invalid or misleading legal expressions (Aryo Pradipta Gema et al., 2025; Ribeiro et al., 2022).

Table 1 compares existing professional-oriented benchmarks and our Pub-LawBench across tasks and dimensions. In experiments, we comprehensively evaluated 17 representative LLMs on Pub-LawBench using user-friendly methods (e.g., simple prompts, Chain-of-Thought), excluding complex setups like agents and RAG. Our experiments reveal current models' limited ability to provide effective public-oriented legal assistance, underscoring the need for more user-centric development and benchmarking. Our experiments reveal current models' limited ability to provide effective public-oriented legal assistance, underscoring the need for more user-centric development and benchmarking.

Our contributions are as follows: (1) We propose Pub-LawBench, a public-oriented benchmark grounded in legal functionalism and genre analysis for lay users. (2) We design three public-oriented evaluation dimensions, i.e., *legal normativity*, *content relevance*, and *format usability*, to complement existing professional-oriented metrics and better serve lay users. (3) We conduct a comprehensive evaluation of 17 LLMs, spanning commercial, open-source, and legal-specialized models, shedding light on their capabilities and shortcomings in public-oriented legal applications.

2 Related Work

Benchmarks on LegalAI. Existing benchmark efforts primarily focus on three major categories: standardized exam benchmarks, e.g., C-

Eval (Huang et al., 2023) and UBE (Katz et al., 2024), which effectively measure legal memorization capabilities but do not fully capture model performance in complex, open-ended legal practice; comprehensive multi-task benchmarks, e.g., LexGLUE (Chalkidis et al., 2022), LegalBench (Guha et al., 2023), LawBench (Fei et al., 2024), LegalAgentBench (Li et al., 2025), which emphasize the assessment of reasoning abilities beyond simple recall; and specialized capability benchmarks, e.g., LeCaRD (Ma et al., 2023), DISC-LawLLM (Yue et al., 2024), and ProvBench (Shen et al., 2025), which targets specific functional requirements(Yu et al., 2026). However, they are largely professional-centric and lack a public-oriented perspective.

Legal Functionalism and Genre Analysis. Public-oriented legal needs are grounded in *legal functionalism* (Pound, 1910; Llewellyn, 1930), which places the value of law in its practical social effects rather than in abstract doctrinal coherence(Mao et al., 2024). This theoretical foundation motivates our design of a public-oriented benchmark tailored to the needs of lay users. Complementing this perspective, *the genre analysis* of legal texts (Bhatia, 1993; Swales, 1990) provides a methodological lens to examine how legal discourse is structured and functionally deployed in specific communicative contexts. For example, a public legal advisory issued by a government agency prioritizes clarity, plain language, and procedural guidance, while court judgments rely on technical jargon and adversarial framing. Together, these insights shape our design of public-oriented tasks and metrics for LLM evaluation, offering valuable insights and support for both research and practice.

3 Pub-LawBench

The core idea of Pub-LawBench is to shift the focus from abstract reasoning depth to the actual utility of LLMs in real-world scenarios. In this section, we will detail two core components of Pub-LawBench: the task taxonomy designed for public scenarios and the evaluation metrics.

3.1 Task Taxonomy

We categorize public legal demands into *Instant QA* and *Legal Gen* based on the user's objectives and required output form. The distinction lies in whether users seek immediate answers or structured documents for legal action.

Instant QA. This category addresses the need for quick, conversational consultation. We classify tasks as Instant QA when the user starts with an informal inquiry about a specific life situation. Unlike professional legal research, which relies on structured case files and dense statutes, public-oriented QA must handle vague or emotional language. In this setting, the model acts as an advisor that simplifies the law for users. We structure these tasks into three levels. *(1) Legal Understanding:* This assesses whether models can identify key facts in a user's story. It consists of JEC-QA-KD (Zhong et al., 2019), CAIL (Xiao et al., 2018), Unfair-ToS (Lippi et al., 2019), and LEdGAR (Tuggener et al., 2020). *(2) Legal Reasoning:* This tests the ability to apply rules to those facts. Tasks include CaseHOLD (Zheng et al., 2021), COLIEE (Rabelo et al., 2020), and JEC-QA-CA (Zhong et al., 2019). *(3) Legal Expression:* This module uses QA-Corpus and Laws-QA (Liu, 2019) to assess the delivery of concise, relevant legal advice without redundancy, ensuring the advice is clear and concise. We detail task definitions in Appendix B.1.

Legal Gen. This category focuses on the production of functional legal materials. Our classification standard for Legal Gen is the creation of documents that carry legal consequences. While professional legal generation often involves drafting complex litigation files or reports, the public needs tools to exercise their rights directly in daily life. Drawing on Genre Analysis (Bhatia, 1993), we evaluate four drafting tasks that represent the typical stages of legal document preparation. *Fact Extraction (Fact)* requires the model to organize various case details into a consistent description. *Reasoning (Reas)* tests the ability to connect those facts to logical conclusions. *Defense (Def)* focuses on identifying valid arguments to support the user's position, while *Judgment Generation (Judg)* evaluates how models apply evidence to reach a final decision. This module ensures that outputs are practically ready for non-experts by prioritizing content relevance, legal normativity, and format usability. This focus on real-world utility distinguishes our approach from purely professional evaluations. Detailed scoring metrics are provided in Appendix B.2.

Layer	Dataset	Lang	Size	Task Description
Instant QA	CAIL	Zh	100k	Fact extraction & Judgment
	JEC-QA-KD	Zh	26k	Concept & Definition QA
	COLIEE	En	5k	Case retrieval & Entailment
	CaseHOLD	En	53k	Holding application
	JEC-QA-CA	Zh	26k	Element subsumption
	Unfair-ToS	En	5k	Unfair terms detection
	LEdGAR	En	80k	Provision classification
	Laws-QA	Zh	20k	Statute interpretation
	QA-Corpus	Zh	35k	Open-ended Q&A
Legal Gen	CaseGen-Fact	Zh	15k	Fact summarization
	CaseGen-Reas	Zh	15k	Reasoning generation
	CaseGen-Judg	Zh	15k	Verdict generation
	CaseGen-Def	Zh	10k	Defense arguments

Table 2: Overview of datasets in **Pub-LawBench**. **Instant QA** covers retrieval and classification, while **Legal Gen** focuses on document drafting tasks.

Data Collection. To ensure holistic coverage across cognitive dimensions, we integrate heterogeneous corpora and Common Law systems. As detailed in Table 2, sources include judicial records, professional exams, and compliance documents. This diversity effectively mitigates single-task bias. We standardized all data into (instruction, input, output) triples, employing the public-oriented prompt templates illustrated below.

▶ Role Setting: You shall act as a professional legal assistant/senior lawyer.

▶ Style Setting: Your responses must adhere to a rigorous, accurate, and professional legal language style.

▶ General Constraint: Analyses and responses are only based on the provided materials. [Objective Multiple-Choice / Classification] Provide the final answer directly.

▶ Conditional Constraint: [Legal Document Drafting] Follow legal document structure strictly, and embody 3 dimensions: Factual: Clarify facts, refute false accusations. Legal: Cite relevant laws. Claim: Clearly put forward claims.

3.2 Evaluation Metrics

In Pub-LawBench, we first evaluate model performance using the original metrics defined by each legal task. These are referred to as *performance-oriented metrics* in this work and primarily measure factual or accuracy. Specifically, for *Instant QA*, we adopt standard classification metrics, such as accuracy, precision, and recall. The official evaluation metric for each subtask is detailed in Appendix B.1. For Legal Gen, we assess the quality of generated content through widely used BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and LLM-as-a-Judge (Zheng et al., 2023).

Then, we introduce *public-oriented metrics* to evaluate whether the model outputs are direct, actionable, and practically useful for lay users. We design the following three evaluation dimensions.

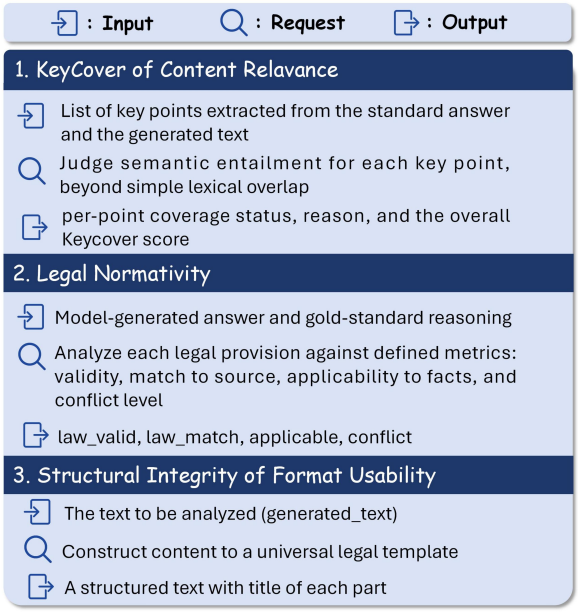

Figure 2: Structured prompt templates for three public-oriented metrics.

3.2.1 Content Relevance

Legal content generated for lay users should include all essential legal key points. Even the omission of seemingly minor information can cause confusion or misunderstanding among non-expert users. To assess comprehensiveness, we propose *Key-point Coverage*, denoted as KeyCover (S_{Key}). We implement it automatically by adopting a GPT-4o powered "extraction–matching" paradigm (Liu et al., 2023), where we first extract atomic key points $K = \{k_1, \ldots, k_n\}$ from reference R and verify their semantic entailment in generation G:

$$S_{Key} = \frac{1}{n} \sum_{i=1}^{n} I(k_i \in G), \tag{1}$$

where $\mathbb{I}$ is the indicator function. The used prompt template is shown in Fig. 2.

3.2.2 Legal Normativity

This metric assesses legal reliability. Given the high stakes of legal advice, we prioritize substantive accuracy over linguistic style (Ji et al., 2023). We aggregate *Statutory Citation*, *Logical Consistency*, and *Terminological Professionalism* with weights assigned in a $2:2:1$ ratio.

To validate our 2:2:1 weighting ratio, we conducted a sensitivity analysis comparing it against

alternative configurations (e.g., 1:1:1, 3:1:1). Our setting achieved the highest Spearman correlation ($\rho = 0.94$) with expert judgment (detailed results in table 9).

Statutory Citation Accuracy To mitigate hallucinations, we employ a dual-verification mechanism. The score combines validity and applicability:

$$S_{Citation} = \text{Validity} \times \text{Applicability}, \quad (2)$$

where validity involves strict database matching, and applicability uses GPT-4o (OpenAI, 2023) to verify contextual relevance. The used prompt template is shown in Fig. 2.

Logical Consistency Using Chain-of-Thought (CoT) (Wei et al., 2022), we detect factual and reasoning contradictions, penalizing conflicts:

$$S_{Logic} = 1 - \frac{w_1 \cdot C_{fact} + w_2 \cdot C_{reason}}{T_{sentences}}. \quad (3)$$

Lower conflict rates indicate higher logical rigor.

Terminology Professionalism We quantify lexical richness via term density (D_{term}) and diversity (V_{term}) based on THUOCL (Han and Li, 2016), calculating the proportion and count of distinct professional terms.

$$D_{term} = \frac{|T_{legal}|}{N}, \quad (4)$$

$$V_{term} = \frac{|\text{unique}(T_{legal})|}{|T_{legal}| + \epsilon}, \quad (5)$$

where $|\cdot|$ denotes the count, unique$(\cdot)$ represents the set of distinct terms, and ϵ is a smoothing term.

3.2.3 Format Usability

We assess practical judicial value by integrating structural integrity (S_{Struct}) and semantic similarity (S_{Sem}). This is a key consideration for public use, as practical judicial value depends on both correct legal structure and faithful semantic content. S_{Struct} is computed by parsing the text into a template-based tree. Additionally, S_{Sem} measures cosine similarity of segments using Lawformer (Xiao et al., 2021), preventing masked local flaws. The final score is:

$$S_{Format} = S_{Struct} + \text{Average}(S_{Sem}). \quad (6)$$

Task	Metric	GPT-4o	Human	Corr.
Fact	Keyword	2.14	1.85	0.93
Reas	Keyword	1.75	1.52	0.93
	Normativity	2.89	2.50	0.94
	Format	2.62	2.25	0.92
Judg	Keyword	1.94	1.68	0.93
	Normativity	2.98	2.58	0.94
	Format	1.84	1.59	0.92
Def	Keyword	2.35	2.03	0.93
	Normativity	3.00	2.60	0.94
	Format	1.84	1.60	0.92

Table 3: **Meta-Evaluation of GPT-4o as a Judge.** We report the Pearson correlation (Corr.) between GPT-4o automated scores and human expert ratings across varying tasks and dimensions. The high correlation (> 0.92) validates the reliability of using GPT-4o for scalable evaluation.

4 Experiments

4.1 Experimental Setup

We evaluate 17 representative LLMs across three categories: (1) *Commercial*: GPT-4 Turbo (OpenAI, 2023), Claude Sonnet 4.5 (Anthropic, 2024), Gemini 2.5 Flash/3 (Gemini Team, Google, 2025; Pichai et al., 2025), DeepSeek-V3/R1 (DeepSeek-AI, 2025, 2024), Qwen3-Max (Bai et al., 2023), Kimi k1.5 (Moonshot AI, 2024); (2) *Legal Specialized*: DISC-LawLLM (Yue et al., 2024), Chatlaw (Cui et al., 2023), LawGPT (Nguyen et al., 2023), SaulLM-7B (Colombo et al., 2024); and (3) *General Open-Weights*: Qwen2.5-3B/7B (Bai et al., 2023), Llama 3.2 3B/3.1 8B (Dubey et al., 2024). More details about baselines are provided in Appendix C.1. We utilize nucleus sampling with $p = 0.9$ and temperature $T = 0.2$. Further deployment details are provided in Appendix C.2.

We use GPT-4o as an automatic judge and validate its effectiveness by comparing its ratings with human assessments on a total of 100 randomly selected responses across all models. Across 5 independent runs, standard deviations remained consistently below 0.003, confirming that the observed performance gaps reflect intrinsic model capabilities rather than random noise (see Appendix D for details). As shown in Table 3, GPT-4o exhibits a high correlation (>0.92) with human experts, confirming its reliability.

Model	Instant QA									Legal Gen				Avg.
	CAIL (Acc.)	JEC-K (Acc.)	COL (BS)	Case (Acc.)	JEC-C (Acc.)	ToS (Acc.)	Ldg (Acc.)	Law (BS)	QA (BS)	Fact (BS)	Reas (BS)	Judg (BS)	Def (BS)	
Commercial LLMs														
DeepSeek-V3	**54.3**	**72.7**	79.5	81.0	70.0	58.0	52.5	82.7	83.8	76.2	71.8	76.5	74.4	**71.8**
Gemini 3	40.9	71.0	82.6	89.0	**76.9**	**86.3**	**75.0**	29.4	55.6	77.0	**73.1**	80.3	72.5	70.0
Claude Sonnet 4.5	42.2	57.5	82.6	**91.9**	53.5	59.2	43.9	82.9	82.8	77.9	72.8	77.0	73.3	69.0
Gemini 2.5 Flash	**54.3**	60.5	81.8	82.0	50.5	42.3	41.3	83.6	84.0	74.8	71.6	75.8	73.3	67.4
GPT-5.2	47.7	54.0	81.8	81.5	56.0	81.4	58.2	28.7	56.8	**78.0**	71.3	79.2	72.3	65.2
Qwen3-Max	53.8	55.0	**86.3**	75.6	42.7	62.0	5.5	83.5	**84.4**	76.5	71.3	74.2	**75.4**	65.1
GPT-4 Turbo	46.2	44.0	80.7	83.0	38.5	56.0	22.0	**83.7**	**84.4**	73.7	71.5	71.0	71.4	63.6
Kimi k1.5	33.7	64.0	78.1	80.0	41.7	44.0	4.0	83.0	83.6	75.1	71.6	72.7	74.1	62.0
DeepSeek-R1	48.8	51.8	79.5	74.0	49.0	34.3	15.0	28.5	56.6	68.8	69.7	67.5	67.1	54.7
General Open-Weights LLMs														
Qwen2.5-7B	34.7	10.9	81.0	68.0	10.8	2.0	1.8	56.4	56.4	64.7	64.1	70.4	72.7	45.7
Qwen2.5-3B	36.2	9.6	81.2	59.5	10.7	2.0	1.0	56.4	56.4	64.7	64.0	69.4	72.6	44.9
Llama 3.1 8B	36.7	9.8	81.5	6.0	11.0	2.0	2.5	54.6	54.7	65.1	63.8	59.2	72.8	40.0
Llama 3.2 3B	36.4	9.2	77.8	1.0	10.7	2.0	2.3	45.1	45.2	64.3	63.4	59.2	72.3	37.6
Legal Specialized LLMs														
SaulLM-7B	5.0	13.8	83.1	65.0	17.6	6.1	1.2	57.2	56.9	65.0	65.4	**84.1**	64.2	45.0
DISC-LawLLM	29.8	7.8	71.0	21.5	10.9	2.0	1.0	58.5	58.6	63.6	63.0	63.3	72.2	40.2
Chatlaw	11.1	11.8	81.1	15.0	7.2	3.0	0.8	58.3	58.4	65.1	63.5	61.6	72.8	39.1
LawGPT	10.8	9.5	83.9	2.5	8.3	4.0	1.4	44.8	47.0	64.3	63.4	59.2	72.3	36.3

Table 4: Overall performance of various LLMs on Pub-LawBench across performance-oriented metrics. *Acc.* denotes accuracy; *BS* denotes BERTScore. Heatmap colors indicate relative performance (darker is better).

4.2 Performance-oriented Comparison

Table 4 summarizes overall model performance on Pub-LawBench across key performance-oriented metrics (e.g., Accuracy and BERTScore). Full results (including F1, Precision, and Recall) are provided in Appendix E. We observed:

In overall performance, Commercial LLMs significantly outperform both small General Open-Weights LLMs and Legal Specialized LLMs. Based on overall average scores (Avg.), Commercial LLMs like DeepSeek-V3 (71.8) and Gemini 3 (70.0) show better performance. However, average scores for General Open-Weights LLMs generally fall below 46, while Legal Specialized LLMs cluster within the 36–45 range. This indicates a correlation between model scale and the ability to handle complex public legal tasks.

In legal understanding and reasoning tasks, model capabilities exhibit distinct task-specific performance tiers. Commercial LLMs demonstrate stable performance on simpler, comprehension-oriented tasks (e.g., COL, Law), but their capabilities decline in high-intensity legal reasoning tasks (e.g., CAIL, ToS), resulting in diminished credibility for complex assignments. Meanwhile, General Open-Weights LLMs and Legal Specialized LLMs exhibit more pronounced performance declines in complex tasks, with most scoring below 20 and some approaching random levels. This indicates that complex legal reasoning tasks impose higher demands on model scale and reasoning capabilities.

Model	Def	Judg	Reas	Fact	Avg.
Commercial LLMs					
DeepSeek-V3	81.8	58.2	49.9	69.8	64.9
Gemini 3	**85.1**	73.6	**89.2**	**93.4**	**85.3**
GPT-5.2	82.6	77.3	85.1	90.6	83.9
Qwen3-Max	82.0	58.2	75.7	59.4	68.8
GPT-4 Turbo	77.9	56.2	76.2	69.9	70.0
DeepSeek-R1	82.2	76.5	86.1	89.8	83.7
General Open-Weights LLMs					
Qwen2.5-7B	84.2	70.6	85.7	86.9	81.8
Qwen2.5-3B	83.9	68.6	84.8	86.8	81.0
Llama 3.1 8B	79.6	39.5	77.4	80.8	69.3
Llama 3.2 3B	81.4	53.3	79.6	81.2	73.9
Legal Specialized LLMs					
Chatlaw	82.4	36.7	81.2	85.7	71.5
SaulLM-7B	77.5	**89.7**	78.1	77.8	80.8

Table 5: **Public-oriented evaluation** on *Content Relevance* using the KeyCover score (%) for across four legal generation tasks. Heatmap colors indicate performance (darker is better). The complete results are provided in Appendix E.

The localized advantages of Legal Specialized LLMs in specific subtasks cannot be translated into stable performance across tasks and scenarios. A few Legal Specialized LLMs (e.g., SaulLM-

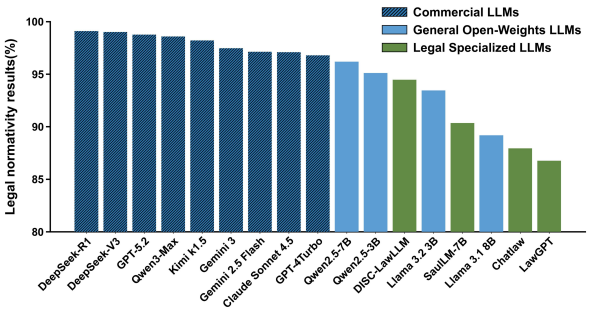

Figure 3: **Public-oriented evaluation** on *Legal Normativity*. The results summarize the models' ability to adhere to statutory standards and avoid hallucinations.

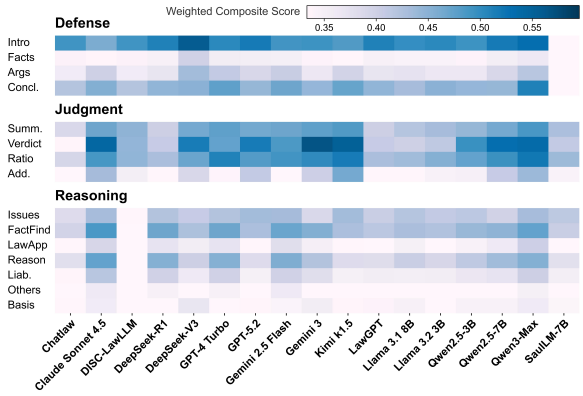

Figure 4: **Public-oriented evaluation** on *Format Usability Score*. Heatmap analysis of the overall model performance on across different tasks.

7B) achieve high scores on the Def task but exhibit limited performance on other tasks, lowering their average capabilities. This phenomenon indicates that current models remain unstable in comprehensive legal scenarios. Their outputs are better used as supplementary references and should not be directly treated as admissible legal conclusions.

4.3 Public-oriented Comparison

We further report the performance on our designed public-oriented metrics across all LLMs on LegalGen, including *content relevance*, *legal normativity*, and *format usability*.

Content Relevance Table 5 reports KeyCover scores that measure how effectively models retain essential legal facts. Most commercial models and Qwen2.5 above 0.8, indicating a strong ability to preserve core public concerns when transforming informal narratives into legally reasoned outputs. However, high fact capture does not guarantee logical consistency. As shown in Appendix Table 15, NLI scores are consistently lower than KeyCover, indicating that models often resort to keyword stuffing rather than sound reasoning. For the public, this misleading relevance is more dangerous than complete irrelevance because it appears helpful while being legally unsound. Notably, SaulLM-7B achieves the highest score on the Judg task (0.897) despite being trained exclusively on English data, suggesting that fundamental legal reasoning patterns may generalize across languages. However, several legal specialized models score below 0.5, reflecting a failure to align specialized legal knowledge with non-professional user expression, and underscoring that in public-facing legal services, legal expertise alone is ineffective without grounding in user intent and language.

Legal Normativity Fig. 3 shows that Commercial models dominate this dimension, consistently achieving scores above 0.97. And General Open-Weight models outperform several Legal Specialized models in this metric. This suggests that, for public legal assistance, a strong general foundation with stable reasoning behavior may be safer than narrowly specialized models that lack strict logical control. Although Legal Specialized models demonstrate familiarity with legal terminology, their comparatively lower scores indicate a heightened risk of legal hallucinations. Indeed, a combined analysis reveals a significant gap between relevance and normativity. While many models achieve high KeyCover scores in they often struggle with legal normativity. This contrast proves that professional knowledge does not equal normative logic. For public-oriented applications, relevance ensures that the model understands the user's problem, but normativity ensures the advice is safe to follow. It is essential for the general public because a relevant but self-contradictory response can lead to incorrect legal actions.

Format Usability Fig. 4 visualizes the Structure Availability Score, reflecting how effectively models organize legal documents. Most models perform well in producing standardized document structures, particularly in clearly delineating surface-level sections such as the Facts component in the Def task. However, performance drops sharply in the Judg task when deeper structural components are required. Although commercial models can generate explicit verdict statements, nearly all models fail to fully populate the addendum (Add.) and reasoning sections. This exposes a key weakness for public-facing legal assistance:

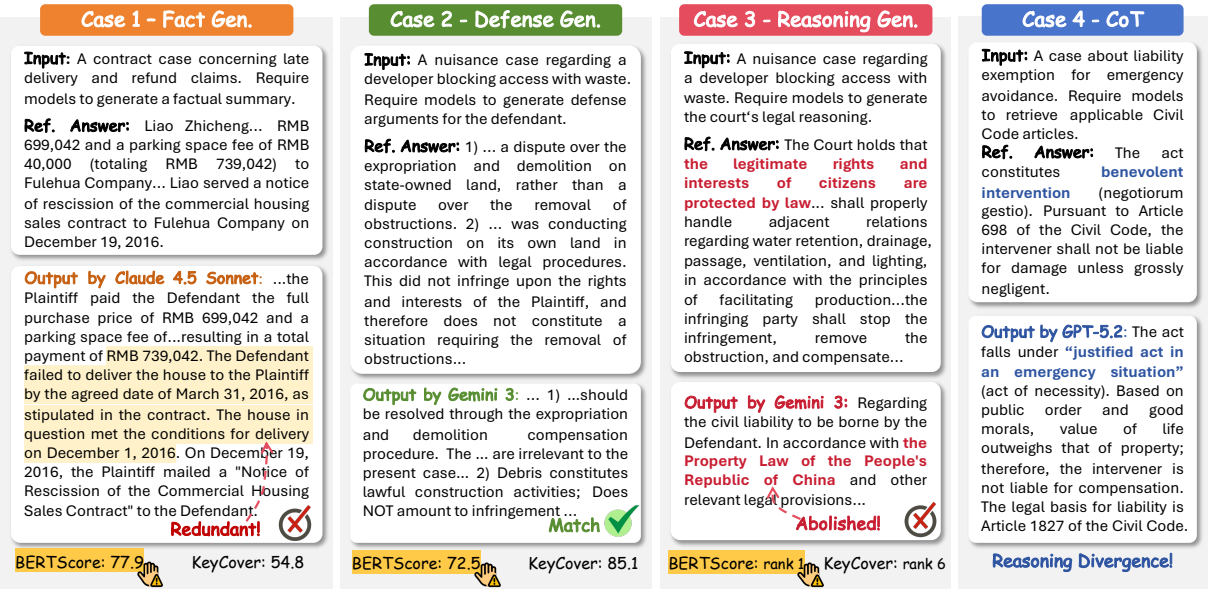

Figure 5: Case Study. Case 1-3 show the output of different LLMs using simple prompting. Metrics highlighted in yellow are misleading, e.g., incorrectly penalizing correct answers or rewarding wrong ones. Case 4 shows the output of CoT prompting, which introduces noise that leads to reasoning divergence rather than enhanced clarity.

Task	Model	Std.	CoT	Δ
Casehold *(Accuracy)*	DeepSeek-V3	81.0	77.0	-4.0
	Claude Sonnet 4.5	91.9	84.9	-7.0
	Gemini 3	89.0	87.7	-1.3
CAIL *(Accuracy)*	DeepSeek-V3	54.3	69.8	+15.6
	Claude Sonnet 4.5	42.2	70.9	+28.7
	Gemini 3	40.9	83.8	+42.9
LEdGAR *(Accuracy)*	DeepSeek-V3	52.5	75.8	+23.3
	Claude Sonnet 4.5	43.9	83.3	+39.4
	Gemini 3	75.0	86.6	+11.6
Unfair-ToS *(Accuracy)*	DeepSeek-V3	58.0	63.6	+5.6
	Claude Sonnet 4.5	59.2	65.7	+6.5
	Gemini 3	86.3	68.7	-17.6
COLIEE *(Rouge-L)*	DeepSeek-V3	19.6	17.2	-2.4
	Claude Sonnet 4.5	21.4	14.7	-6.7
	Gemini 3	21.4	17.7	-3.7
CaseGen-F *(BERTScore)*	DeepSeek-V3	77.0	78.3	+1.3
	Claude Sonnet 4.5	76.2	76.9	+0.7
	Gemini 3	77.9	79.6	+1.7

Table 6: Impact of CoT prompting. Δ denotes the performance gain/loss compared to standard prompting.

models are capable of generating documents that look professionally structured, but often cannot supply the concrete guidance or transparent chains of legal reasoning that users need to understand decisions and take subsequent legal actions.

4.4 In-depth Analysis

Case Study Fig. 5 shows real cases and the generated output of LLMs. It is observed that current LLMs are somehow prone to misleading the public. For instance, in Case 1, the output generated by Claude Sonnet 4.5 contains excessive irrelevant details, resulting in verbose outputs that fail to strictly adhere to the concise nature of legal drafting. Our public-oriented metric, KeyCover (54.8), effectively captures this deficiency, whereas the performance-oriented metric BERTScore (77.9) proves invalid, as it assigns a high score to this flawed output, thereby failing to reflect its inadequacy. It is noteworthy that a misleading metric can also be harmful. In the defendant generation task (Case 2), Gemini-3.1 produced a concise output that accurately captured the core legal arguments for a defense. While our KeyCover metric appropriately validates this correct response, BERTScore assigned it an unjustifiably low score, exposing its inadequacy in assessing legal correctness. Case 3 then reveals a common type of hallucination in LLMs: the cited *Property Law of the People's Republic of China* was abolished, a fact unlikely to be known to the general public. Our Legal Normativity metric sensitively detects this critical error, as evidenced by a sharp drop. These findings reveal that current LLMs still have room for improvement in public-facing legal capabilities.

Impact of CoT Prompting. Contrary to general findings in NLP, our results in Table 6 show that CoT does not consistently benefit. While effective for extraction, CoT hindered performance in reasoning tasks (e.g., CaseHOLD, COLIEE). We observe that for general-purpose or non-expert models, CoT induces a "validity illusion", generating plausible-sounding but legally unsound arguments (as in Case 4 of Fig 5). In multiple-choice constraints, this manifests as reasoning drift, where the model diverges from the core legal issue during

the chain generation, leading to incorrect option selection. This suggests that without specialized post-training, CoT may amplify hallucinations.

Model	Zero-Shot	RAG	Gain
GPT-4o	0.422	0.511	+0.089
Qwen2.5-72B	0.408	0.495	+0.087
DISC-LawLLM	0.245	0.356	+0.111
ChatGLM3-6B	0.198	0.284	+0.086

Table 7: Performance comparison on Law_QA with and without RAG. Integration of RAG consistently enhances performance across models.

Impact of RAG. We evaluated representative models on Law_QA using a BM25-retriever with the National Laws and Regulations Database. Table 7 shows that RAG consistently improves performance (e.g., +11.1 for DISC-LawLLM) without altering the relative ranking of models, demonstrating that our zero-shot findings are indicative of model performance even in RAG-integrated settings.

5 Conclusion

In this work, we proposed **Pub-LawBench** to evaluate the utility of LLMs in public legal services. Our experiments yield four core findings. First, commercial models lead in accuracy, while open-weight models suit basic consultations. Second, legal logic is cross-lingually universal, yet specialized models often fail to align expertise with public needs, causing logic conflicts. Third, models master document frameworks but lack actionable details. Finally, CoT prompting can be counterproductive by inducing reasoning bias in strict contexts. Thus, professional rigor does not automatically ensure public utility. Future work should prioritize output readability and practical guidance for non-experts.

Limitations

While Pub-LawBench offers a practical way to evaluate LegalAI, it has certain boundaries. First, the benchmark focuses on core tasks that represent common public legal needs. It does not yet cover all specialized legal niches or highly rare case types. Second, although our current version includes both Chinese and English datasets, it has not yet been tested on low-resource or minority languages. Legal data in these languages is often scarce, which makes it harder for models to process professional terminology accurately. Finally, more work is needed to adapt the framework to other jurisdictions with different procedural rules. These points provide a clear roadmap for the future growth of the benchmark.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. 63261068), the National Natural Sinence Foundation of China (Grant No. 72571150) and the Academy for Advanced Interdisciplinary Studies, Nankai University (AAIS-NKU-2025-21).

Ethic Statements

All datasets used in this study are open-source to ensure transparency and encourage further research. To protect privacy, we applied strict anonymization to all legal documents by removing personal names, identification numbers, and other sensitive identifiers. Despite these precautions, our findings show that current models still produce hallucinations and logical errors. Therefore, we strongly advise against using any existing legal LLMs as a formal basis for official legal decisions or documentation. These systems are intended as research tools and should not be treated as a substitute for professional legal advice or authoritative evidence.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Anthropic Model Card*.

Ahmed Abdulaal Tom Diethe Philip Alexander Teare Beatrice Alex Pasquale Minervini Amrutha Saseendran Aryo Pradipta Gema, Chen Jin and 1 others. 2025. Decore: Decoding by contrasting retrieval heads to mitigate hallucinations. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP)*, page 10003–10039.

John Langshaw Austin. 1962. *How to Do Things with Words*. Oxford University Press.

Jinze Bai, Shuai Bai, Shusheng Yang, and 1 others. 2023. Qwen-vl: A frontier of large multimodal models. *arXiv preprint arXiv:2308.12966*.

Vijay K Bhatia. 1993. *Analysing Genre: Language Use in Professional Settings*. Longman.

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

Odysseas S. Chlapanis, Dimitris Galanis, Nikolaos Aletras, and Ion Androutsopoulos. 2025. GreekBarBench: A challenging benchmark for free-text legal reasoning and citations. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP)*, pages 25099–25119.

Pierre Colombo, Telmo Piguet, Boudin Hui, and 1 others. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

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

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 641(7978):583–590.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, and 1 others. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7933–7962.

Gemini Team, Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Re, Adam Chilton, Arvind Narayana, Alex Chien, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xu Han and Maosong Li. 2016. Thuocl: Tsinghua university open chinese lexicon.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Prathamesh Kalamkar, Astha Agarwal, Abhinav Tiwari, Smita Gupta, Saurabh Karn, Vivek Patil, and Ashutosh Modi. 2023. SemEval-2023 task 6: LegalEval - understanding legal texts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval)*, pages 2211–2223.

Hidetaka Kamigaito, Jingun Kwon, and Manabu Okumura. 2023. Abstractive document summarization with summary-length prediction. In *Findings of the Association for Computational Linguistics: European Chapter of the Association for Computational Linguistics (Findings of EACL)*, pages 606–612.

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).

Jihyung Lee, DaeHee Kim, Seonjeong Hwang, Hyounghun Kim, and Gary Lee. 2025. KoBLEX: Open legal question answering with multi-hop reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out (ACL Workshop)*, pages 74–81.

Marco Lippi, Przemyslaw Palka, Giuseppe Contissa, and 1 others. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139.

Huanyong Liu. 2019. Crimekgassitant: A chinese legal intelligence project.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Karl N. Llewellyn. 1930. A realistic jurisprudence—the next step. *Columbia Law Review*.

Yixiao Ma, Yiqun Shao, Yueyue Wu, Yihong Liu, Ruizhe Zhang, and Min Zhang. 2023. LeCaRD: A legal case retrieval dataset for chinese law system. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2742–2752.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2342–2348.

Rui Mao, Tianwei Zhang, Qian Liu, Amir Hussain, and Erik Cambria. 2024. Unveiling diplomatic narratives: Analyzing united nations security council debates through metaphorical cognition. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, pages 1–7.

Moonshot AI. 2024. Kimi: A long-context language model. `https://www.moonshot.cn`. Accessed: 2026-04-24.

Ha-Thanh Nguyen, P. M. Nguyen, Vuong M. Vu, and 1 others. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. *arXiv preprint arXiv:2302.05729*.

Shubham Kumar Nigam, Aniket Deroy, Subhankar Maity, and Arnab Bhattacharya. 2024. Rethinking legal judgement prediction in a realistic scenario in the era of large language models. In *Proceedings of the Natural Legal Language Processing Workshop (NLLP)*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Sundar Pichai, Demis Hassabis, Koray Kavukcuoglu, and Gemini Team, Google DeepMind. 2025. A new era of intelligence with Gemini 3. Google Blog. Accessed: 2026-04-24.

Roscoe Pound. 1910. Law in books and law in action. *American Law Review*, 44(1).

Juliano Rabelo, Mi-Young Kim, Yoshinobu Kano, and Randy Goebel. 2020. Overview of the competition on legal information extraction/entailment (coliee) 2020. In *New Frontiers in Artificial Intelligence (JSAI-isAI)*.

Leonardo F Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. Factgraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3234–3251.

Luca Rolshoven, Vishvaksenan Rasiah, Srinanda Brügger Bose, Sarah Hostettler, Lara Burkhalter, Matthias Stürmer, and Joel Niklaus. 2025. Unlocking legal knowledge: A multilingual dataset for judicial summarization in Switzerland. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (Findings of EMNLP)*, pages 15382–15411, Suzhou, China. Association for Computational Linguistics.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Xiuxuan Shen, Zhongyuan Jiang, Junsan Zhang, Junxiao Han, Yao Wan, Chengjie Guo, Bingcheng Liu, Jie Wu, Renxiang Li, and Philip S. Yu. 2025. Provbench: A benchmark of legal provision recommendation for contract auto-reviewing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4578–4590.

Abhay Shukla and 1 others. 2022. Legal case document summarization: Extractive and abstractive methods. In *Proceedings of the 2nd International Workshop on AI and Intelligent Assistance for Legal Professionals (LegalNLP)*.

John M Swales. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. Ledgar: A large-scale multi-label corpus for contract provision classification. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1235–1241.

Qiqi Wang, Guanjin Wang, Yihong Pan, Zhipeng Lin, Huijia Li, Qian Liu, and Kaiqi Zhao. 2025. Husk: A hierarchically structured urban knowledge graph dataset for multi-level spatial tasks. In *Proceedings of the ACM Conference on Data Science and Management of Data (CODASD)*, pages 1–6.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

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

Chaojun Xiao, Lan Xue, Yuan Zhang, Tian Tian, Zhiyuan Liu, and Maosong Sun. 2021. Lawformer:

A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

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. In *Proceedings of the 17th Chinese Computational Linguistics Conference (CCL)*.

Hang Yu, Qiqi Wang, and Qian Liu. 2026. Legal knowledge infusion for large language models: A survey. *Information Fusion*, 107:102390.

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, and 1 others. 2024. Lawllm: Intelligent legal system with legal reasoning and verifiable retrieval. In *International Conference on Database Systems for Advanced Applications (DASFAA)*, pages 304–321.

Tianyi Zhang, Varsha Kishore, Felix Wu, and 1 others. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL)*, pages 159–168.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Jec-qa: A legal-domain question answering dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2771–2780.

A Related work

With the development of LLMs, LegalAI are widely use LLMs into several tasks. Existing methods can be broadly divided into two categories. The first category relies on prompt-based methods to elicit legal reasoning capabilities from general-purpose models by guiding them to follow structured legal logic. The second category introduces domain-specific legal knowledge, either by training or fine-tuning legal LLMs on large-scale legal corpora to access statutes during inference. While these methods have demonstrated effectiveness in improving reasoning performance, they typically require careful prompt design, specialized legal knowledge, or complex system configurations. As a result, such approaches are less accessible to non-specialist audiences, compelling us to evaluate the performance of LLMs with minimal user intervention.

Existing evaluation works primarily focus on three categories: Standardized Exam Benchmarks, Comprehensive Multi-task Benchmarks, and Specialized Capability Benchmarks. Standardized Exam Benchmarks evaluate models against human professional standards. Examples include JEC-QA (Zhong et al., 2019)for Chinese judicial knowledge. While effective for testing legal memorization, they do not fully reflect performance in complex, open-ended legal practice. Comprehensive Multi-task Benchmarks evaluate reasoning beyond simple recall. Key benchmarks include LegalBench (162 tasks focusing on IRAC analysis and interpretation) (Guha et al., 2023), LawBench (a hierarchical system covering memorization, understanding, and application) (Fei et al., 2024), which uses complexity-stratified scenarios to test model limits. Specialized Capability Benchmarks focus on specific functional requirements. LeCaRD (Ma et al., 2023)evaluates legal case retrieval precision; LegalHallucination diagnoses the frequency and types of factual errors; and DISC-LawLLM (Yue et al., 2024)tests legal reasoning within autonomous agent interactions.

While current benchmarks cover a wide range of professional tasks, they are primarily focused on professionals and lack benchmarks grounded in the public perspective. Specifically, we aim to evaluate LLMs output under legal normativity, the structural integrity of documents, and the appropriateness of generated formats for official use. Our work aims to bridge this gap by assessing the actual utility of LegalAI in public service.

B Definitions of our evaluation tasks

B.1 Instant QA

Fact Extraction and Judgment Prediction This task (Xiao et al., 2018) provides a case description containing mixed or misleading information, requiring the model—like a judge—to accurately extract key legal facts and, in accordance with criminal law provisions, predict the correct charges and sentencing. This tests the model’s ability to “separate truth from falsehood” and “make decisions based on the law”. Performance

CAIL: Fact Extraction and Judgment Prediction

问题: 关于毒品犯罪，下列哪些选项正确?

A. 已满 14 周岁不满 16 周岁的未成年人参与毒品的制造、运输、贩卖的，只对贩卖行为承担刑事责任。

B. 以自己吸食为目的非法持有毒品的，一律不构成犯罪。

C. 对不满 18 周岁的未成年人不适用毒品犯罪特别再犯制度，是对行为人有利的类推解释。

D. 非法种植毒品原植物，在收获前自动铲除的，成立犯罪中止。

Answer: A

Question: Regarding drug-related crimes, which of the following statements are correct?

A. If a minor aged between 14 and 16 participates in the manufacturing, transportation, or sale of drugs, criminal liability shall be imposed only for the act of sale.

B. Illegal possession of drugs solely for personal consumption does not, in all cases, constitute a criminal offense.

C. The non-application of the special recidivism regime for drug-related crimes to minors under the age of 18 constitute an analogy favorable to the offender.

D. Illegal cultivation of drug-source plants, where the plants are voluntarily destroyed before harvest, constitutes an attempted offense.

Answer: A

Figure 6: A sample in CAIL for *Fact Extraction and Judgment Prediction* task. The upper is the original text, and the lower is the English translation.

on this task is evaluated using accuracy, precision, and recall.

Concept and Definition QA This task (Zhong et al., 2019) formulates questions regarding abstract legal terminology and provides highly misleading answer options. This assesses whether the model possesses precise knowledge of legal dictionaries and can distinguish between everyday language and rigorous legal terminology. Performance on this task is evaluated using accuracy, precision, and recall.

Case Retrieval and Entailment This task (Rabelo et al., 2020) provides a scenario from daily life, requiring the model to bridge the semantic gap, retrieve the corresponding civil law provisions, and perform logical reasoning to determine whether compensation is required under those provisions. This assesses the model's ability to map real-life facts to legal rules through logical deduction. Performance on this task is evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1.

JEC-CA: Concept and Definition QA

问题: 如何理解我国《宪法》规定："中华人民共和国公民在法律面前一律平等"?

A. 任何公民都受法律约束，不允许有违法特权。

B. 所有公民在司法上平等，在法律的实施、执行和适用上平等。

C. 适用法律上平等。

D. 在法律面前一律平等是指公民权利能力上的平等，而不是行为能力上的平等。

E. 在法律面前一律平等是指法律范围内的平等，而不是事实上的平等。

Answer: C

Question: How should the constitutional principle that "all citizens of the People's Republic of China are equal before the law" be understood?

A. All citizens are bound by the law, and no unlawful privileges are permitted.

B. All citizens are equal in the judiciary and in the enactment, enforcement, and application of the law.

C. Equality before the law refers to equality in the application of the law.

D. Equality before the law refers to equality in legal capacity, not equality in capacity for civil conduct.

E. Equality before the law refers to equality within the scope of the law, not equality in fact.

Answer: C

Figure 7: A sample in JEC-CA for *Concept and Definition QA* task. The upper is the original text, and the lower is the English translation.

COLIEE: Case Retrieval and Entailment

Query Case

In cases where an individual rescues another person from getting hit by a car by pushing that person out of the way, causing the person's luxury kimono to get dirty, the rescuer does not have to compensate for damages for the kimono.

Retrieved Statute

Article 698 (Civil Code): "If a manager engages in benevolent intervention in another's business in order to allow a principal to escape imminent danger to the principal's person, reputation, or property, the manager is not liable to compensate for damage resulting from this unless the manager has acted in bad faith or with gross negligence."

Figure 8: A sample in COLIEE for *Case Retrieval and Entailment* task. The upper is the question, and the lower is the reference answer.

Holding Application and Analogy This task (Zheng et al., 2021) provides a case context that

CaseHOLD: Holding Application and Analogy

Case Description

"...have passed into the hands of bona fide creditors or purchasers for value, as long as any debts of the corporation are unpaid, the holders of the assets take them charged with a trust in favor of the creditors."
Rice v. City of Columbia, 143 S.C. 516, 141 S.E. 705, 712 (1928).
"[T]he very moment a corporation, banking or other, reaches the point of insolvency...its assets become impressed with a solemn trust to be distributed ratably among its creditors...and the managing agents become the administrators of that trust."
Stewart v. Ficken, 151 S.C. 424, 149 S.E. 164, 165 (1929).

Candidate Holdings

A. Recognizing the right of a creditor to sue a corporate director for breach of fiduciary duty.

B. Holding that a plaintiff seeking individual relief under ERISA...

C. Holding that breach of fiduciary duty claim was preempted by FEHBA.

D. Holding a cause of action for breach of fiduciary duty will not lie where...

E. Holding that Missouri law applied...

Answer: A

Figure 9: A sample in CaseHOLD for *Holding Application and Analogy* task. The upper is the original text, and the lower is the reference answer.

cites a specific legal principle, and requires the model to identify the correct judgment conclusion for another similar case from among multiple options. This does not test rote memorization, but rather assesses the model's ability to understand the internal logical consistency of judicial precedents and apply them analogically. Performance on this task is evaluated using accuracy, precision, and recall.

Element Subsumption This task (Zhong et al., 2019) sets up a case with complex details, requiring the model to compare these fragmented facts one by one with the constitutive elements of various criminal offenses (subsumption). This tests the model's refined qualitative ability to judge whether a given behavior satisfies all statutory conditions of a particular crime. Performance on this task is evaluated using accuracy, precision, and recall.

Unfair Terms Detection This task (Lippi et al., 2019) provides a lengthy terms of service agreement containing clauses that infringe on consumer rights, such as "unilateral modification at any time

JEC-KD: Element Subsumption

问题: 张某从人贩子那里买了杨某，当时并不知道杨某还没有满 14 岁，从而与杨某发生了性关系，事后知道了，就没有再与之发生性关系，直到杨某满 14 岁后才再次同居。而杨某也愿意跟从张某生活，在公安人员来解救时，张某与杨某四处躲藏，则张某的行为应该构成的罪名有:

A. 强奸罪

B. 收买被拐卖的妇女、儿童罪

C. 奸淫幼女罪

D. 阻碍解救被拐卖妇女儿童罪

Answer: BD

Question: Zhang bought Yang from a human trafficker without initially knowing that Yang was under 14 years old, and had sexual intercourse with Yang under these circumstances. After learning the truth, Zhang ceased such acts and only resumed cohabitation after Yang turned 14. Yang was willing to live with Zhang. When police arrived to rescue Yang, Zhang and Yang hid to evade them. The offenses Zhang may be charged with include:

A. Rape.

B. Crime of purchasing abducted women or children.

C. Crime of sexual intercourse with a minor under 14.

D. Crime of obstructing the rescue of abducted women or children.

Answer: BD

Figure 10: A sample in JEC-KD for *Element Subsumption* task. The upper is the original text, and the lower is the English translation.

LEdGAR & Unfair-TOS

LEdGAR: Provision Classification

"this amendment and any claims, controversy, dispute or causes of action (whether in contract or tort or otherwise) based upon, arising out of or relating to this amendment shall be construed in accordance with and governed by the laws of the state of New York."
Answer: [4] Governing Law / Choice of Law

Unfair-TOS: Unfair Terms Detection

"academia.edu reserves the right, at its sole discretion, to modify the site, services and these terms, at any time and without prior notice."
Answer: [2] Unilateral Change

Figure 11: The upper part is a sample in LEdGAR for *Provision Classification* task, and the lower part is a sample in unfair-tos for *Unfair Terms Detection* task.

without notice" and "mandatory arbitration." This tests the model's ability to identify legally unequal risk points between rights and obligations, serving the function of compliance review. Performance on this task is evaluated using accuracy, precision, and recall.

LawsQA: Statute Interpretation

问题: “你好律师，如果男方不同意打胎女方执意打胎法律有何规定?”
Answer: 男方的生育权是受到限制的，女方有权利是否决定流产，男方不能阻止女方流产。理由是，女方怀孕后，胎儿成为女方身体的一部分，是否引产属于其人身权利的范围，男方的生育权不可能高于女方的人身自由权……首先，我国婚姻法中并没有生育权的规定。其次，《妇女权益保障法》规定，妇女有按照国家有关规定生育子女的权利，也有不生育的自由。因此，女方既有生育的权利，也有不生育的自由。

Question: “Hello lawyer, if the man does not agree to an abortion but the woman insists, what are the relevant legal provisions?”
Answer: The man's reproductive rights are limited. The woman has the right to decide whether to have an abortion, and the man cannot prevent it. The reason is that after conception, the fetus becomes part of the woman's body, and the decision to terminate the pregnancy falls within the scope of her personal rights. The man's reproductive rights cannot override the woman's right to personal freedom. First, China's Marriage Law does not stipulate reproductive rights. Second, the Law on the Protection of Rights and Interests of Women stipulates that women have the right to bear children in accordance with state regulations, as well as the freedom not to bear children. Therefore, the woman has both the right to bear children and the freedom not to bear children.

Figure 12: A sample in LawsQA for *Statute Interpretation* task. The upper is the original text, and the lower is the English translation.

QA-Corpus: Open-ended QA

问题: “手部伤残鉴定七级大概能赔偿多少钱?”
Answer: 二十万

Question: “How much is the approximate compensation for a Grade 7 hand disability assessment?”
Answer: Two hundred thousand

Figure 13: A sample in QA-Corpus for *Open-ended QA* task. The left is the original text, and the right is the English translation.

Provision Classification This task (Tuggener et al., 2020) presents a section of a contract written entirely in uppercase complex legal text, and requires the model to classify it into standard clause categories based on its semantic features. This tests the model's understanding of the functions of legal text and its ability to process structure. Performance on this task is evaluated using accuracy, precision, and recall.

Statute Interpretation This task (Liu, 2019) requires the model to comprehensively apply multiple statutes to provide a systematic interpretation and balance of legal interests, confronted with a complex issue not directly regulated by law. This tests the model's advanced interpretative capability in addressing legal gaps and conflicts of rights. Performance on this task is evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1.

Open-ended QA This task (Liu, 2019) presents a real-world, challenging, and ambiguous dilemma, with no preset answer options, requiring the model to provide step-by-step, actionable, practical advice as a lawyer would. This tests the model's comprehensive consulting ability to solve actual problems. Performance on this task is evaluated using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1.

B.2 Legal Gen

To address the limitations of traditional n-gram metrics in capturing the semantic depth and logical rigor of legal texts, we established a comprehensive evaluation framework comprising three core dimensions: Content Relevance, Legal Normativity, and Format Usability. This appendix details the definitions and calculation methodologies for specific metrics within each dimension.

B.2.1 Content Relevance

Content Relevance assesses whether the model accurately addresses the user's legal intent and covers critical case facts. This dimension relies on two complementary metrics: Key Point Coverage and Intent Alignment Score.

To quantify the completeness of information, we introduce S_{Key}. Unlike fuzzy semantic matching, this metric employs a rigid “extraction-matching” protocol using GPT-4. Initially, the judge model extracts a set of atomic key information points, denoted as $K = \{k_1, k_2, ..., k_n\}$, from the ground truth reference R. Subsequently, the generated text G is evaluated against K to verify whether each atomic point k_i is semantically entailed. The final score is calculated as $S_{Key} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(k_i \in G)$, where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if the key point is present, and 0 otherwise.

B.2.2 Legal Normativity

Legal Normativity evaluates the model's adherence to authoritative legal standards through three sub-dimensions: Statutory Citation Accuracy,

Logical Consistency, and Terminological Professionalism.

Addressing the hallucination issue in legal LLMs, we design a dual verification mechanism for Statutory Citation Accuracy. First, a Legal Validity Check extracts all statute names and articles using regular expressions and cross-references them with the National Laws and Regulations Database to calculate the validity ratio (N_{valid}/N_{total}). Second, a Statutory Applicability Check leverages GPT-4 to analyze the case context, identifying valid but misapplied statutes to determine the applicability ratio ($N_{relevant}/N_{valid}$). The comprehensive citation score is the product of these two ratios: $S_{Citation}$ = Validity $\times$ Applicability.

For S_{Logic}, we assess the rigor of legal argumentation. Using a specialized CoT prompt, the judge model detects Factual Contradictions (inconsistencies with input facts) and Reasoning Contradictions (disconnects between premises and conclusions). The score is derived by penalizing these errors relative to text length: $S_{Logic} = 1 - \frac{w_1 \cdot C_{fact} + w_2 \cdot C_{reason}}{T_{sentences}}$, where C_{fact} and C_{reason} represent error counts weighted by w_1 and w_2, respectively.

Finally, Terminological Professionalism quantifies the mastery of legal vocabulary. Using the THUOCL (Han and Li, 2016) as a benchmark, we calculate D_{term}, defined as the proportion of matched legal terms to the total word count, alongside the Unique Term Count to reflect lexical diversity.

B.2.3 Format Usability

Format Usability evaluates the practical utility of the output for judicial practice, focusing on strict adherence to document schemas.

This dimension integrates S_{Struct} and S_{Sem}. For standardized documents like civil judgments, we define a gold standard schema. A structural parser analyzes the generated text tree to identify missing or misplaced sections, calculating S_{Struct} based on structural integrity penalties. To prevent global averaging from masking local defects in long texts, we employ a segmentation mechanism. The generated text G and reference R are segmented into corresponding sections (e.g., Plaintiff's Argument, Court's Opinion), and the cosine similarity between their vector representations is calculated using a legal-specific pre-trained model. Here we use Lawformer (Xiao et al., 2021) as the model. The final score is a weighted sum: $S_{Format} = S_{Struct} + \text{Average}(S_{Sem})$, ensuring both structural correctness and local semantic accuracy are captured.

C Experiment setup details

C.1 Baselines

To rigorously evaluate the boundaries of LLM capabilities in the legal domain, we conduct experiments on a comprehensive set of 17 models. We categorize these baselines into three distinct groups to investigate the performance gap between closed-source SOTA and open-source alternatives, as well as the trade-off between general reasoning and domain-specific knowledge.

Commercial SOTA Models We evaluate 9 commercial models representing the pinnacle of current AI capabilities, serving as the topline performance benchmarks.

ChatGPT-5.2: Widely regarded as the current state-of-the-art LLM. Compared to its predecessors, ChatGPT-5.2 demonstrates a qualitative leap in complex logical reasoning and long-horizon planning, setting the absolute ceiling for legal reasoning tasks.

ChatGPT-4-Turbo: The industrial standard for stability and instruction following, serving as a robust reference point for evaluating model reliability. (OpenAI, 2023)

Claude Sonnet 4.5: Renowned for its safety alignment and nuanced understanding of complex text. Its inclusion allows us to assess performance in legal scenarios where precise wording and intent alignment are critical. (Anthropic, 2024)

Gemini 3 & Gemini 2.5 Flash: Representing the evolution of Google's long-context capabilities. Gemini 3, in particular, excels in processing massive legal corpora within a single context window. (Gemini Team, Google, 2025; Pichai et al., 2025)

DeepSeek-V3: A leading Chinese MoE (Mixture-of-Experts) model, demonstrating performance comparable to top-tier global models in Chinese logic and knowledge tasks. (DeepSeek-AI, 2024)

DeepSeek-R1: A specialized reasoning-intensive model optimized via Reinforcement Learning (RL). It is included to investigate the potential of RL-driven Chain-of-Thought (CoT) in solving multi-step judicial deduction problems. (DeepSeek-AI, 2025)

Qwen3-Max: The flagship model of the Qwen series, representing the highest standard of Chinese commercial models in language understanding and generation. (Bai et al., 2023)

Kimi k1.5: Distinguished by its lossless long-context recall, making it highly effective for tasks involving extensive statute retrieval and case history analysis. (Moonshot AI, 2024)

Domain-Specific Open-Source Models This group consists of 4 models explicitly fine-tuned or pre-trained on legal corpora to test the efficacy of domain knowledge injection.

DISC-LawLLM-13B: Built upon the Baichuan-13B backbone. It utilizes a high-quality dataset covering legal reasoning and judicial examinations for Supervised Fine-Tuning (SFT), representing a robust baseline for larger-scale open legal LLMs. (Yue et al., 2024)

Chatlaw-7B: Based on the InternLM-7B architecture. This model incorporates external knowledge-base retrieval mechanisms and is fully fine-tuned on Chinese statutes, excelling at precise statute citation. (Cui et al., 2023)

LawGPT-7B: An early representative based on the Chinese-LLaMA (Alpaca) backbone. Fine-tuned on legal QA datasets, it serves as a baseline for early-stage instruction-tuned legal models. (Nguyen et al., 2023)

SaulLM-7B: Based on Mistral-7B but underwent continual pre-training on massive English legal corpora. It serves as a cross-jurisdiction baseline to evaluate performance differences between Civil Law and Common Law systems. (Colombo et al., 2024)

General-Purpose Open-Source Models We selected 4 general-purpose base models to investigate the transferability of general reasoning skills to the legal domain and the feasibility of edge deployment.

Qwen2.5-7B: Currently the strongest base model in the 7B class. Its superior logic generalization, derived from extensive math and code training, tests whether general reasoning can outperform domain-specific fine-tuning. (Bai et al., 2023)

Qwen2.5-3B: A lightweight variant of the Qwen series, included to assess the minimal parameter threshold required for basic legal reasoning. (Bai et al., 2023)

Llama 3.1 8B: Meta's standard open-source baseline. With 8B parameters, it serves as a benchmark for evaluating the cross-domain adaptability of Western general models on Chinese legal tasks. (Dubey et al., 2024)

Llama 3.2 3B: Optimized for edge devices. We evaluate this model to determine the feasibility of deploying privacy-preserving legal assistants on resource-constrained hardware. (Dubey et al., 2024)

C.2 Implementation Details

C.2.1 Local Model Deployment and Adaptation

All open-source models were evaluated on a cluster of 8 NVIDIA RTX 3090 (24GB) GPUs. To ensure numerical consistency, we standardized inference in FP16 half-precision. For the larger DISC-LawLLM-13B, we employed Tensor Parallelism across dual GPUs to avoid precision degradation associated with 8-bit quantization. To address compatibility bottlenecks arising from heterogeneous architectures (e.g., Llama 3's RoPE scaling and Baichuan's tokenizer), we implemented a unified adaptation layer with configuration remapping, ensuring robust inference across diverse model generations.

C.2.2 API Model Configuration

Commercial models (e.g., GPT-5.2, DeepSeek-V3) were accessed via their official APIs. To maintain a fair comparison with local baselines, we aligned the decoding hyperparameters (Temperature=0.2, Top-p=0.9) and disabled external tools (e.g., web browsing) to evaluate intrinsic model capabilities. For reasoning-intensive models like DeepSeek-R1, we parsed their Chain-of-Thought (CoT) outputs separately from the final answer to analyze reasoning depth.

Model	Total Score (Std)	Legal Norm. (Std)	Content Rel. (Std)	Format Usability (Std)
GPT-4o	0.0012	0.0021	0.0018	0.0015
Qwen2.5-72B	0.0015	0.0025	0.0022	0.0019
DISC-LawLLM	0.0022	0.0028	0.0026	0.0021
ChatGLM3-6B	0.0025	0.0030	0.0029	0.0024
Average	**< 0.0020**	**< 0.0030**	**< 0.0030**	**< 0.0025**

Table 8: Statistical stability analysis (Standard Deviation) across five independent runs.

Weight Combination	Spearman ρ vs. Human Expert
1:1:1	0.85
3:1:1	0.89
1:3:1	0.88
1:1:3	0.81
2:2:1 (Ours)	**0.94**

Table 9: Sensitivity analysis of the weight distribution in Legal Normativity.

Prompt Variant	Variant 1	Variant 2	Variant 3
Variant 1	1.00	0.96	0.93
Variant 2	0.96	1.00	0.97
Variant 3	0.93	0.97	1.00

Table 10: Spearman rank correlation coefficients (ρ) between different prompt variants.

C.2.3 Inference Setup

For generative tasks, we utilized nucleus sampling (p=0.9) with a low temperature (0.2) and a repetition penalty (1.1) to balance fluency with factual rigor. For discriminative tasks, greedy decoding was employed. We strictly aligned prompt templates (e.g., ChatML, Alpaca) with each model's pre-training distribution and implemented a stateful checkpointing mechanism to handle long-context generation stability.

D Robustness and Stability Analysis

This section provides supplementary analyzes to validate the reliability and objectivity of our framework.

D.1 Sensitivity Analysis of Legal Normativity Weights

The *Legal Normativity* metric in is composed of three sub-dimensions: citation accuracy, logical consistency, and terminology professionalism. To justify the 2:2:1 weight ratio used in our main experiments, we conducted a sensitivity analysis by comparing our configuration against several alternative weighting schemes.

As shown in Table 9, we calculated the Spearman rank correlation coefficient (ρ) between the model rankings produced by each scheme and the gold-standard rankings provided by human legal experts. The results indicate that the 2:2:1 ratio achieves the highest alignment with expert judgment ($\rho = 0.94$), confirming that our weighting effectively prioritizes substantive legal accuracy.

D.2 Robustness to Prompt Variations

To ensure that the benchmark results are not overly sensitive to specific prompt engineering, we designed three variants of evaluation instructions. *Variant 1* (our default) provides detailed scoring rubrics; *Variant 2* uses a simplified, concise instruction format; and *Variant 3* requires the model to provide a brief rationale before outputting a score.

Table 10 presents the Spearman correlation between model rankings obtained under these variants. The high correlation across all pairs ($\rho > 0.93$) demonstrates that is robust to reasonable variations in prompt formulation, ensuring that the performance gaps we observe reflect intrinsic model capabilities.

D.3 Statistical Stability and Variance Analysis

Given the potential non-determinism in LLM outputs, we assessed the statistical stability of our benchmark by conducting five independent runs for each baseline model. We report the standard deviation (Std) for the total score and each evaluation dimension.

Table 8 shows that the variance across runs is consistently low (average Std < 0.003). This low variance confirms that the performance rankings reported in our main text are statistically significant and not the result of random noise during inference.

E Detailed Experimental Results

This appendix presents the comprehensive quantitative evaluation results for all 17 models. To provide a holistic view of model capabilities, we report data across six detailed tables, categorizing performance by model type and evaluation dimension.

Comprehensive Performance (Tables 11 & 12) These two tables summarize the fundamental performance of **Massive LLMs** and **Open-Weights/Specialized LLMs** across all 13 tasks. They report Accuracy for understanding tasks (e.g., Casehold, JEC-QA) and KeyCover scores for generation tasks, serving as the baseline for assessing general legal competency.

Legal Normativity and Terminology (Tables 13 & 14) Table 13 evaluates the models' adherence to legal standards, specifically measuring:

- *Validity & Citation Match*: The rate of authentic vs. hallucinated statute citations.
- *Usability & Logic Conflict*: The practical applicability and logical consistency of the generated legal reasoning.

Complementing this, Table 14 analyzes **Terminological Professionalism**. It details the *Density*, *Diversity*, and *Stability* of legal vocabulary used by models, reflecting their "lawyer-like" linguistic style.

Content Semantic Quality (Table 15) While KeyCover measures information recall, Table 15 provides a deeper semantic analysis of the generated text using advanced metrics:

- *Rouge-L & BERTScore*: Assessing n-gram overlap and semantic similarity against expert-written references.
- *NLI (Natural Language Inference)*: Verifying whether the model's output is logically entailed by the ground truth, ensuring factual correctness.

Structural Integrity (Table 16) Table 16 breaks down the **Format Usability** score into specific document sections (e.g., *Introduction*, *Verdict*, *Reasoning*). This fine-grained view highlights specific structural strengths and weaknesses, such as whether a model can correctly format a complex "Verdict" section compared to a simpler "Introduction".

Task	Metric	DeepSeek-V3	Gemini 3	Claude Sonnet 4.5	Gemini 2.5 Flash	Kimi k1.5	GPT-5.2	Qwen3-Max	GPT-4 Turbo	DeepSeek-R1
Casehold	Acc.	81.0	89.0	**91.9**	82.0	80.0	81.5	75.6	83.0	74.0
	F1	81.0	89.1	**91.9**	82.0	80.0	81.5	75.6	83.0	74.1
	Prec.	81.0	89.3	**91.9**	82.0	80.0	82.2	75.6	83.0	75.9
	Rec.	81.0	89.0	**91.9**	82.0	80.0	81.8	75.6	83.0	74.9
COLIEE	R-1	29.9	34.3	34.3	33.0	**45.4**	-	44.5	33.2	29.9
	R-2	8.8	**11.6**	**11.6**	11.3	8.5	11.3	8.9	10.7	8.8
	R-L	19.6	21.4	21.4	20.6	**34.8**	20.6	**34.8**	20.1	19.6
	BS-F1	79.5	82.6	82.6	81.8	**87.3**	81.8	86.3	80.7	79.5
JEC-QA-CA	Acc.	70.0	**76.9**	53.5	50.5	41.7	56.0	42.7	38.5	49.0
	F1	**87.5**	83.0	53.5	50.5	60.6	80.6	65.8	38.5	71.7
	Prec.	**88.4**	83.7	53.5	50.5	60.9	82.0	65.0	38.5	73.2
	Rec.	**90.0**	83.0	53.5	50.5	63.1	82.6	70.8	38.5	73.6
JEC-QA-KD	Acc.	**72.7**	71.0	57.5	60.5	64.0	54.0	55.0	44.0	51.8
	F1	**87.1**	83.6	57.5	60.5	80.1	79.5	77.9	44.0	77.9
	Prec.	**87.7**	84.1	57.5	60.5	79.4	81.2	78.1	44.0	78.0
	Rec.	**88.9**	85.8	57.5	60.5	83.2	82.7	83.6	44.0	82.7
CAIL	Acc.	**54.3**	40.9	42.2	**54.3**	33.7	47.7	53.8	46.2	48.8
	F1	**57.8**	40.9	42.2	54.3	42.8	47.7	56.2	46.2	48.8
	Prec.	**56.8**	40.9	42.2	54.3	39.9	47.7	55.4	46.2	48.8
	Rec.	**60.3**	40.9	42.2	54.3	52.3	47.7	58.3	46.2	48.8
LEdGAR	Acc.	52.5	**75.0**	43.9	41.3	4.0	58.2	5.5	22.0	15.0
	F1	38.0	**56.0**	30.6	29.1	1.2	38.3	2.0	15.5	8.1
	Prec.	39.5	**55.4**	31.7	30.6	0.9	37.7	2.1	15.7	9.0
	Rec.	40.9	**57.7**	32.3	31.1	2.8	41.3	2.0	18.2	8.4
Unfair-ToS	Acc.	58.0	**86.3**	59.2	42.3	44.0	81.4	62.0	56.0	34.3
	F1	**31.7**	28.8	14.6	10.7	15.3	27.4	22.3	16.8	20.4
	Prec.	**28.9**	23.2	13.6	10.1	14.0	22.0	19.8	14.1	17.2
	Rec.	**61.6**	50.0	28.2	23.4	41.6	50.0	41.8	36.4	50.0
Laws-QA	R-1	0.1	-	0.4	0.2	0.3	-	**0.7**	0.4	-
	BS-F1	82.7	-	82.9	83.6	83.0	-	83.5	**83.7**	-
QA-Corpus	R-1	3.5	-	**6.0**	3.7	5.8	-	1.8	2.7	-
	BS-F1	83.8	-	82.8	84.0	83.6	-	**84.4**	**84.4**	-
Fact	KeyCover	76.2	77.0	**77.9**	74.8	75.1	**78.0**	76.5	73.7	68.8
Reas	KeyCover	71.8	**73.1**	72.8	71.6	71.6	71.3	71.3	71.5	69.7
Judg	KeyCover	76.5	**80.3**	77.0	75.8	72.7	79.2	74.2	71.0	67.5
Def	KeyCover	74.4	72.5	73.3	73.3	74.1	72.3	**75.4**	71.4	67.1

Table 11: Comprehensive results for Commercial LLMs. Metrics include Accuracy (Acc.), F1-Score, Precision (Prec.), Recall (Rec.), Rouge (R-1/2/L), BERTScore F1 (BS-F1), and KeyCover.

Task	Metric	General Open-Weights				Legal Specialized			
		Qwen2.5-7B	Qwen2.5-3B	Llama 3.1 8B	Llama 3.2 3B	SaulLM-7B	DISC-LawLLM	Chatlaw	LawGPT
Casehold	Acc.	**68.0**	59.5	6.0	1.0	65.0	-	-	2.5
	F1	**67.9**	59.3	7.0	1.8	64.9	-	-	4.7
	Prec.	68.2	62.6	18.8	22.9	**72.1**	-	-	45.0
	Rec.	**68.4**	59.8	5.8	1.0	66.3	-	-	2.5
COLIEE	R-1	23.3	21.5	26.6	17.1	28.9	28.2	13.2	**33.1**
	R-L	14.5	13.3	16.0	10.5	17.6	17.0	7.1	**21.2**
	BS-F1	81.0	81.2	81.5	77.8	83.1	71.0	81.1	**83.9**
JEC-QA-CA	Acc.	10.8	10.7	11.0	10.7	**17.6**	10.9	7.2	0.0
	F1	11.7	12.0	12.0	11.6	23.4	11.9	**30.7**	11.8
	Prec.	8.2	8.6	8.6	8.1	43.8	8.5	**44.1**	37.0
	Rec.	20.0	20.0	20.0	20.0	22.2	20.0	**24.1**	7.8
JEC-QA-KD	Acc.	10.9	9.6	9.8	9.2	**13.8**	7.8	11.8	0.5
	F1	13.0	12.8	12.8	12.6	24.9	12.7	**39.6**	7.8
	Prec.	9.6	9.4	9.4	9.2	44.3	9.3	**51.9**	18.6
	Rec.	20.0	20.0	20.0	20.0	21.9	20.0	**32.3**	5.1
CAIL	Acc.	34.7	36.2	**36.7**	11.1	5.0	29.8	11.1	11.1
	F1	**36.9**	35.8	13.5	24.2	6.4	15.5	24.2	24.2
	Prec.	34.6	**37.0**	13.4	30.1	21.0	32.9	30.1	30.1
	Rec.	**43.5**	36.8	18.2	20.3	4.1	16.4	20.3	20.3
LEdGAR	Acc.	1.0	-	-	-	1.0	**2.5**	0.0	**2.5**
	F1	**0.9**	-	-	-	0.1	**0.9**	0.0	**0.9**
	Prec.	0.6	-	-	-	0.1	**0.7**	0.0	**0.7**
	Rec.	1.3	-	-	-	1.2	**1.8**	0.0	**1.8**
Unfair-ToS	Acc.	2.0	2.0	2.0	2.0	**6.1**	2.0	3.0	0.0
	F1	0.5	0.5	0.5	0.5	**12.1**	0.5	0.5	0.6
	Prec.	0.2	0.2	0.2	0.2	**7.1**	0.2	0.3	0.3
	Rec.	12.5	12.5	12.5	12.5	**50.0**	12.5	12.5	12.5
Laws-QA	R-1	3.5	3.5	**6.0**	4.7	4.1	1.6	0.5	-
	BS-F1	56.4	56.4	54.6	45.1	57.2	**58.5**	**58.3**	44.8
QA-Corpus	R-1	3.5	3.4	**6.1**	4.0	2.7	1.4	0.2	-
	BS-F1	56.4	56.4	54.7	45.2	56.9	**58.6**	**58.4**	47.0
Fact	KeyCover	64.7	64.7	**65.1**	64.3	65.0	63.6	**65.1**	64.3
Reas	KeyCover	64.1	64.0	63.8	63.4	**65.4**	63.0	63.5	63.4
Judg	KeyCover	70.4	69.4	59.2	59.2	**84.1**	63.3	61.6	59.2
Def	KeyCover	**72.7**	**72.6**	**72.8**	**72.3**	64.2	**72.2**	**72.8**	**72.3**

Table 12: Combined results for General Open-Weights and Legal Specialized LLMs.

Model	Validity				Citation Match				Usability				Logic Conflict			
	D	J	R	F	D	J	R	F	D	J	R	F	D	J	R	F
Commercial LLMs																
Claude Sonnet 4.5	0.980	0.997	0.997	-	0.905	0.996	0.977	-	0.887	0.983	0.943	-	0.095	0.010	0.034	-
Gemini 2.5 Flash	0.997	0.999	0.998	-	0.907	0.989	0.962	-	0.925	0.958	0.962	-	0.089	0.015	0.030	-
GPT-4 Turbo	0.996	0.999	0.998	-	0.903	0.987	0.955	-	0.898	0.963	0.957	-	0.101	0.017	0.039	-
Qwen3-Max	0.999	**1.00**	0.995	-	0.947	**1.00**	0.990	-	0.965	0.973	0.968	-	0.046	0.011	0.016	-
DeepSeek-V3	**1.00**	**1.00**	0.972	-	**0.998**	0.999	0.964	-	**0.997**	0.982	0.946	-	**0.003**	**0.002**	**0.008**	-
Kimi k1.5	0.996	0.997	**1.00**	-	0.928	0.995	**0.997**	-	0.947	0.967	0.950	-	0.058	0.015	0.022	-
DeepSeek-R1	**1.00**	**1.00**	0.986	-	**0.998**	0.999	0.968	-	**0.997**	0.982	0.918	-	**0.003**	**0.002**	0.030	-
Gemini 3	0.994	**1.00**	0.997	-	0.912	0.999	0.977	-	0.905	**0.989**	**0.986**	-	0.116	0.012	0.010	-
GPT-5.2	**1.00**	**1.00**	**1.00**	-	**0.998**	0.999	0.928	-	**0.997**	0.982	0.971	-	**0.003**	**0.002**	0.031	-
General Open-Source LLMs																
Qwen2.5-3B	0.962	0.993	0.995	-	0.887	0.974	0.927	-	0.915	0.955	0.868	-	0.091	0.019	0.099	-
Qwen2.5-7B	0.970	0.996	0.985	-	0.904	0.990	0.931	-	0.906	0.984	0.932	-	0.048	0.011	0.064	-
Llama 3.2 3B	0.981	0.952	0.976	-	0.885	0.927	0.902	-	0.930	0.936	0.889	-	0.089	0.038	0.101	-
Llama 3.1 8B	0.962	0.844	0.973	-	0.885	0.770	0.898	-	0.891	0.750	0.928	-	0.102	0.106	0.077	-
Legal Specialized LLMs																
Chatlaw	0.958	0.840	0.975	-	0.901	0.721	0.879	-	0.909	0.764	0.877	-	0.091	**0.140**	**0.126**	-
DISC-LawLLM	0.980	0.971	0.983	-	0.904	0.937	0.906	-	0.924	0.947	0.885	-	0.070	0.030	0.091	-
LawGPT	0.968	0.711	0.977	-	0.890	0.662	0.908	-	0.909	0.676	0.951	-	**0.117**	0.046	0.055	-
SaulLM-7B	0.968	0.885	0.955	-	0.902	0.833	0.891	-	0.921	0.825	0.863	-	0.084	0.103	**0.129**	-

Table 13: Evaluation of Legal Normativity

Model	Avg Density				Avg Diversity				Avg Count				Max Density				Stability			
	D	J	R	F	D	J	R	F	D	J	R	F	D	J	R	F	D	J	R	F
Commercial LLMs																				
Claude Sonnet 4.5	1.32	2.46	1.69	1.05	8.57	10.60	8.41	4.88	17.35	20.79	15.79	7.71	3.83	4.96	4.68	4.11	0.536	0.710	**0.865**	0.684
Gemini 2.5 Flash	1.36	2.12	1.82	1.19	10.05	9.77	11.23	6.62	25.72	21.34	31.50	12.16	3.26	4.72	4.23	3.11	0.623	0.671	0.756	0.542
GPT-4 Turbo	1.39	2.12	1.78	1.23	9.38	8.82	9.17	5.79	22.93	19.25	23.57	10.01	4.96	5.16	4.88	3.23	0.592	0.759	0.743	0.638
Qwen3-Max	1.41	2.34	1.80	0.945	11.14	13.38	14.02	5.83	25.43	25.70	30.43	9.84	2.83	4.60	4.17	3.39	0.506	0.735	0.670	0.640
DeepSeek-V3	1.22	2.18	1.04	1.02	15.79	**16.02**	10.88	8.44	33.95	35.95	26.62	16.74	2.55	4.32	2.95	2.29	0.413	0.801	0.859	0.456
Kimi k1.5	1.46	2.63	1.75	1.21	8.75	11.71	9.80	6.08	19.34	22.77	20.15	10.31	2.91	5.12	3.74	3.94	0.655	0.888	0.785	0.769
DeepSeek-R1	1.52	0.615	0.509	0.662	5.42	3.38	10.99	4.89	14.15	8.52	35.75	11.00	2.97	4.15	3.41	2.59	0.685	**1.10**	0.769	0.638
Gemini 3	1.32	2.35	1.64	1.38	14.14	6.84	11.96	11.96	31.39	12.39	23.54	23.54	3.25	5.79	4.02	4.02	0.502	0.743	0.737	0.737
GPT-5.2	1.28	2.38	1.84	1.40	17.70	7.84	13.64	7.66	**51.88**	13.84	32.42	15.14	3.15	3.69	3.42	2.82	0.426	0.632	0.717	0.656
General Open-Source LLMs																				
Qwen2.5-3B	1.28	1.95	1.29	1.23	15.70	11.14	13.49	14.10	40.36	27.70	37.54	39.28	3.48	4.86	4.41	3.44	0.569	0.860	0.516	0.496
Qwen2.5-7B	1.28	1.86	1.31	1.21	16.18	11.73	13.99	14.16	41.05	27.24	38.88	38.38	3.65	4.97	2.85	3.19	0.551	0.782	0.478	0.461
Llama 3.2 3B	1.23	1.50	1.22	1.15	12.37	9.15	11.14	10.74	31.39	25.46	28.65	27.63	3.38	5.68	5.22	4.54	0.573	0.892	0.524	0.505
Llama 3.1 8B	0.934	1.45	0.951	0.881	11.47	7.31	10.43	10.04	23.64	21.49	22.94	21.88	**10.65**	**41.85**	**6.22**	2.51	0.585	0.875	0.456	0.376
Legal Specialized LLMs																				
Chatlaw	1.21	1.51	1.20	1.17	14.44	8.60	13.96	13.82	31.22	17.02	31.06	31.41	3.28	5.26	2.92	2.68	0.511	0.785	0.445	0.427
DISC-LawLLM	0.279	0.515	0.354	0.320	4.09	7.12	4.41	3.97	10.46	16.41	13.49	12.10	1.21	1.92	1.15	1.09	0.177	0.384	0.183	0.183
LawGPT	0.512	0.515	0.532	0.495	8.97	7.12	8.78	8.18	16.27	16.41	17.20	16.06	1.41	1.92	1.80	1.52	0.229	0.384	0.236	0.231
SaulLM-7B	0.512	0.515	0.532	0.495	8.97	7.12	8.78	8.18	16.27	16.41	17.20	16.06	1.41	1.92	1.80	1.52	0.229	0.384	0.236	0.231

Table 14: Analysis of Legal Words

Model	Rouge-L				KeyCover				NLI				BERTScore			
	D	J	R	F	D	J	R	F	D	J	R	F	D	J	R	F
Commercial LLMs																
Claude Sonnet 4.5	0.340	0.177	0.222	0.236	0.778	0.513	0.646	0.548	0.504	0.489	0.489	0.475	0.733	0.649	0.679	0.668
Gemini 2.5 Flash	0.316	0.235	0.293	0.284	0.797	0.574	**0.780**	0.727	0.518	0.495	0.507	0.478	0.717	0.665	0.691	0.680
GPT-4 Turbo	0.311	0.230	0.275	0.271	0.779	0.562	0.762	0.699	0.517	0.493	0.498	0.483	0.714	0.664	0.691	0.686
Qwen3-Max	**0.353**	0.202	0.301	0.271	0.821	0.583	0.757	0.594	**0.558**	0.490	0.502	0.456	**0.746**	0.659	0.686	0.652
DeepSeek-V3	0.312	0.206	0.161	0.277	0.818	0.582	0.499	0.698	0.497	0.504	0.482	0.450	0.722	0.644	0.608	0.666
Kimi k1.5	0.330	0.189	0.257	0.293	0.745	0.538	0.692	0.648	0.548	0.491	0.505	0.492	0.739	0.663	0.685	0.675
DeepSeek-R1	0.303	**0.358**	0.266	0.346	0.822	**0.765**	**0.861**	**0.898**	0.486	0.488	0.473	0.449	0.671	0.675	0.697	0.688
Gemini 3	0.302	**0.500**	0.288	**0.466**	**0.851**	0.735	**0.892**	**0.934**	0.551	**0.539**	**0.545**	**0.532**	0.725	**0.803**	**0.731**	**0.770**
GPT-5.2	0.261	**0.503**	0.252	0.459	0.826	**0.773**	**0.851**	**0.906**	0.508	0.527	**0.549**	**0.512**	0.723	**0.792**	0.713	**0.780**
General Open-Source LLMs																
Qwen2.5-3B	0.223	0.276	0.223	0.257	0.839	0.686	**0.848**	**0.868**	0.252	0.334	0.137	0.077	0.726	0.694	0.640	0.647
Qwen2.5-7B	0.224	0.284	0.224	0.261	0.842	0.706	**0.857**	**0.869**	0.253	0.343	0.138	0.077	0.727	0.704	0.641	0.647
Llama 3.2 3B	0.236	0.157	0.212	0.269	0.814	0.533	**0.796**	**0.812**	0.515	0.499	**0.552**	0.506	0.729	0.649	0.640	0.651
Llama 3.1 8B	0.242	0.122	0.207	0.272	0.796	0.395	0.774	**0.808**	0.514	0.509	**0.552**	0.506	0.728	0.592	0.638	0.651
Legal Specialized LLMs																
Chatlaw	0.222	0.102	0.194	0.247	0.824	0.367	**0.811**	**0.857**	0.482	0.522	0.537	0.506	0.728	0.616	0.635	0.651
DISC-LawLLM	0.178	0.126	0.162	0.188	0.534	0.402	0.518	0.433	0.534	0.477	0.496	0.469	0.722	0.632	0.630	0.635
LawGPT	0.256	0.199	0.203	0.280	0.770	0.428	0.724	**0.775**	0.500	0.449	0.492	0.456	0.723	0.592	0.634	0.643
SaulLM-7B	0.275	**0.718**	0.289	0.285	0.774	**0.897**	0.781	**0.777**	0.505	**0.542**	0.519	**0.509**	0.642	**0.842**	0.655	0.650

Table 15: Evaluation of Content Relevance

Model	Defense Task				Judgment Task				Reasoning Task						
	Intro	Facts	Args	Concl.	Summ.	Verd.	Ratio	Add.	Iss.	Fact	Law	Reas	Liab	Oth.	Basis
Commercial LLMs															
Kimi k1.5	0.483	0.335	0.396	0.471	0.483	0.552	0.518	**0.464**	**0.434**	0.427	0.335	0.384	0.351	0.290	0.338
Qwen3-Max	0.531	0.364	0.418	**0.512**	**0.487**	0.536	**0.521**	0.439	0.432	0.474	**0.402**	0.453	0.402	0.350	0.359
Gemini 3	0.494	0.341	0.366	0.442	0.476	**0.567**	0.502	0.403	0.389	0.455	0.351	0.422	0.380	0.315	0.355
Gemini 2.5 Flash	0.490	0.353	0.405	0.468	0.468	0.485	0.490	0.334	0.434	0.470	0.381	0.458	0.400	0.362	0.342
Claude Sonnet 4.5	0.458	0.338	0.400	0.455	0.467	0.543	0.488	0.433	0.432	**0.486**	0.391	**0.473**	**0.417**	**0.362**	0.357
DeepSeek-V3	**0.559**	**0.400**	**0.434**	0.450	0.464	0.517	0.470	0.390	0.407	0.426	0.355	0.401	0.361	0.297	**0.370**
DeepSeek-R1	0.512	0.348	0.371	0.446	0.402	0.407	0.428	0.334	0.422	0.468	0.372	0.452	0.399	0.348	0.330
GPT-5.2	0.519	0.355	0.391	0.441	0.465	0.518	0.481	0.410	**0.434**	0.427	0.335	0.384	0.351	0.290	0.338
GPT-4 Turbo	0.504	0.356	0.410	0.476	0.476	0.473	0.510	0.339	0.422	0.468	0.372	0.452	0.399	0.348	0.330
General Open-Source LLMs															
Qwen2.5-7B	0.519	0.355	0.391	0.441	0.464	0.532	0.495	0.407	0.399	0.447	0.361	0.435	0.372	0.314	0.345
Qwen2.5-3B	0.494	0.341	0.366	0.442	0.442	0.496	0.474	0.349	0.415	0.440	0.347	0.411	0.367	0.317	0.346
Llama 3.1 8B	0.498	0.344	0.373	0.432	0.420	0.401	0.435	0.335	0.420	0.428	0.351	0.404	0.355	0.321	0.348
Llama 3.2 3B	0.503	0.344	0.365	0.449	0.433	0.413	0.453	0.346	0.409	0.424	0.327	0.391	0.356	0.310	0.329
Legal Specialized LLMs															
DISC-LawLLM	0.492	0.343	0.358	0.420	0.447	0.444	0.445	0.355	0.336	0.324	0.283	0.316	0.299	0.293	0.305
LawGPT	0.512	0.348	0.371	0.446	0.402	0.407	0.428	0.334	0.404	0.416	0.321	0.382	0.356	0.307	0.341
Chatlaw	0.492	0.343	0.358	0.420	0.389	0.338	0.392	0.313	0.383	0.396	0.335	0.378	0.337	0.307	0.335
SaulLM-7B	0.000	0.003	0.010	0.000	0.427	0.410	0.437	0.344	0.413	0.404	0.328	0.377	0.355	0.301	0.340

Table 16: **Fine-grained Structural Evaluation.** We report the weighted composite scores for each logical section across three tasks. **Bold** indicates the highest score in each column. Note that while general SOTA models maintain high scores (> 0.45) across Judgment sections, specialized models like Chatlaw exhibit significant drops in complex sections like *Verdict* and *Addendum*.