LexGenius: An Expert-Level Benchmark for Large Language Models in Legal General Intelligence

Wenjin Liu[1,2,*] Haoran Luo[2,*] Xin Feng[1] Xiang Ji[1] Lijuan Zhou[1†]
Rui Mao[2] Jiapu Wang [3†] Shirui Pan [4] Erik Cambria[2]
[1]Hainan University [2]Nanyang Technological University
[3]Nanjing University of Science and Technology [4]Griffith University
wenjinliu23@outlook.com, haoran.luo@ieee.org

Homepage GitHub Dataset HF Models

Abstract

Legal general intelligence (GI) refers to artificial intelligence (AI) that encompasses legal understanding, reasoning, and decision-making, simulating the expertise of legal experts across domains. However, existing benchmarks are result-oriented and fail to systematically evaluate the legal intelligence of large language models (LLMs), hindering the development of legal GI. To address this, we propose LexGenius, an expert-level Chinese legal benchmark for evaluating legal GI in LLMs. It follows a Dimension-Task-Ability framework, covering seven dimensions, eleven tasks, and twenty abilities. We use the recent legal cases and exam questions to create multiple-choice questions with a combination of manual and LLM reviews to reduce data leakage risks, ensuring accuracy and reliability through multiple rounds of checks. We evaluate 12 state-of-the-art LLMs using LexGenius and conduct an in-depth analysis. We find significant disparities across legal intelligence abilities for LLMs, with even the best LLMs lagging behind human legal professionals. We believe LexGenius can assess the legal intelligence abilities of LLMs and enhance legal GI development. Our project is available at https://github.com/QwenQKing/LexGenius.

1 Introduction

"The law is the expression of the general will."

— Jean-Jacques Rousseau

Legal general intelligence is the capacity of general AI to perform with expert-level ability across complex legal contexts (e.g., hard tasks, soft intelligence) (Kant et al., 2025; Zhou et al., 2025). It involves the precise interpretation of legal provisions, sound inference based on complex factual scenarios (Zhang et al., 2025b; Li et al., 2025e; Shen et al., 2025), the resolution of conflicts among rules from multiple interrelated legal domains, and the ability to make normatively binding judgments (Yue et al., 2024a; Zhang et al., 2025d; Luo et al., 2025b) in uncertain and ethically sensitive contexts (Kim et al., 2025; Huang et al., 2023; Liu et al., 2025c). Legal general intelligence is not just whether AI knows the law, but whether it can participate in the normative structure of legal systems, thereby opening the door to its integration into legal order.

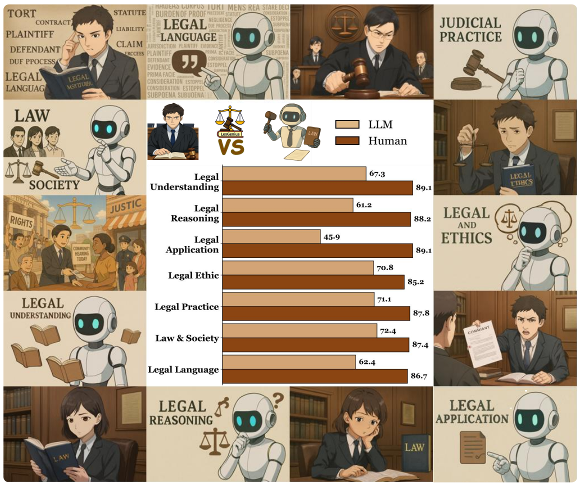


Figure 1: Comparison of the state-of-the-art LLMs and human legal experts illustrates that humans outperform LLMs in the seven legal intelligence dimensions.

In recent years, LLMs have demonstrated strong performance across general language tasks (Mao et al., 2024; Zheng et al., 2025). This progress has catalyzed a surge of interest in adapting LLMs to the legal domain, aiming to tackle challenges, such as legal question-answering (Su et al., 2025; Fei et al., 2024), case analysis (Zhang et al., 2025a; Li et al., 2025a), and judgment prediction (Liu et al., 2025b; Xie et al., 2025). To assess the legal reasoning capabilities of LLMs, several benchmarks, such as LegalBench (Guha et al., 2023), LexEval (Li et al., 2024a), and LexGLUE (Chalkidis et al., 2021; Jia et al., 2025), are introduced. These benchmarks provide a critical foundation for evaluating, improving, and advancing the legal capabilities of large language models (Luo et al., 2025a,c).

* Equal contribution. † Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 18565–18585
July 2-7, 2026

However, the existing benchmarks encounter the following limitations: **(1) Legal intelligence has not yet entered the second half of AI.** Current benchmarks (Chalkidis et al., 2021; Fei et al., 2024) focus on technical tasks while neglecting soft legal intelligence, such as ethical judgment, the law–morality boundary, and societal impact assessment (Wang et al., 2024; Cambria et al., 2024). **(2) Data contamination and lack of comprehensiveness.** Static, publicly available legal benchmarks risk data leakage (Wu et al., 2025) and fail to assess models on dynamic reasoning or novel legal scenarios, leading to overstated and unreliable evaluations. **(3) Lack of a structured framework for comprehensively assessing legal intelligence abilities.** Outcome-focused benchmarks overlook legal reasoning stages, blurring the line between true understanding and pattern mimicry (Cui et al., 2023; Zhang et al., 2026b,a; Chen et al., 2024).

To address the above limitations, we propose LexGenius, a comprehensive benchmark to assess legal general intelligence for LLMs. First, we rethink the evaluation of legal intelligence for LLMs (Thakur et al., 2024). Recognizing that existing benchmarks (Wang et al., 2024; Chang et al., 2024; Xu et al., 2025a; Fei et al., 2024) overlook aspects of legal soft intelligence, our framework explicitly incorporates tests of capabilities such as ethical judgment, moral-legal boundaries, and social impact. We have developed a new collection of 8,385 standardized legal multiple-choice questions (MCQs), covering civil, criminal, and commercial law. To ensure legal accuracy, all questions and answers are refined through professional review. These MCQs assess multi-dimensional competencies (Cambria et al., 2024; Corfmat et al., 2025) relevant to legal intelligence. Furthermore, to address cognitive coverage limitations in existing benchmarks, we propose a framework structured around the dimensions of legal theory and practice, organizing tasks and abilities to reflect real-world legal intelligence. Focusing on Chinese laws ensures a robust and meaningful assessment, as distinct legal systems would otherwise dilute the evaluation.

Leveraging LexGenius, we evaluated 12 state-of-the-art (SOTA) LLMs and 2 prompting strategies, including naive and chain-of-thought (CoT) prompting (Kojima et al., 2022). A baseline performance, constructed by 6 legal professionals, was also established for comparison. Results show that even the top-performing LLM, DeepSeek-R1 (Guo et al., 2025), exhibits a significant gap compared to human experts across various legal general intelligence abilities (Yao et al., 2025b; Hannah et al., 2025; Dong et al., 2025), as shown in Figure 1. In summary, our contributions in this work include:

- We propose the LexGenius, a three-level evaluation framework (Dimension-Task-Ability), for systematically and comprehensively evaluating the legal intelligence capabilities of LLMs.
- We introduce legal soft intelligence into the legal intelligence evaluation of LLMs, paving the way for the assessment of legal general intelligence to move towards the second half of AI.
- We evaluate 12 SOTA LLMs on LexGenius and analyze their gaps and limitations in legal intelligence at different levels and perspectives.

2 Related Work

We review existing benchmarks for LLMs, including legal benchmarks and expert-level benchmarks:

Legal Benchmarks. Recently, a series of legal benchmarks have emerged (see Table 1). They have made significant contributions to evaluating LLMs' performance (Kanapala et al., 2019; Yao et al., 2025a), including retrieval (STARD, LeCaRD) (Su et al., 2024; Li et al., 2024b), question answering (JEC-QA, Legal CQA) (Zhong et al., 2020; Askari et al., 2022), classification (LexGLUE) (Chalkidis et al., 2021), reasoning (LegalBench, LexEval) (Guha et al., 2023; Li et al., 2024a), and others (Laiw and LawBench) (Dai et al., 2025; Fei et al., 2024). However, most benchmarks remain task-oriented and outcome-focused, offering limited insight into the underlying legal general intelligence of LLMs (Yue et al., 2024b).

Expert-level Benchmarks. To usher in the sec-

Benchmark	Lan.	M-Dim.	F-Gra.	Com.	Soft Int.	Ato. Abi.
STARD	CN	✗	✗	✓	✗	✗
LexGLUE	EN	✓	✗	✓	✗	✗
LegalBench	EN	✓	✓	✓	✗	✗
LeCaRDv2	CN	✗	✗	✗	✗	✗
JEC-QA	CN	✓	✗	✓	✗	✗
LawBench	CN	✓	✗	✓	✗	✗
Legal CQA	EN	✗	✗	✓	✗	✗
LexEval	CN	✓	✗	✓	✓	✗
Laiw	CN	✓	✗	✗	✗	✗
Ours	CN	✓	✓	✓	✓	✓

Table 1: Comparison of the existing benchmarks and LexGenius (ours). Lan. means Language; M-Dim. means Multi-Dimensional; F-Gra. means Fine-Grained; Com. means Comprehensiveness; Soft Int. means Soft Intelligence; and Ato. Abi. means Atomicized Ability.

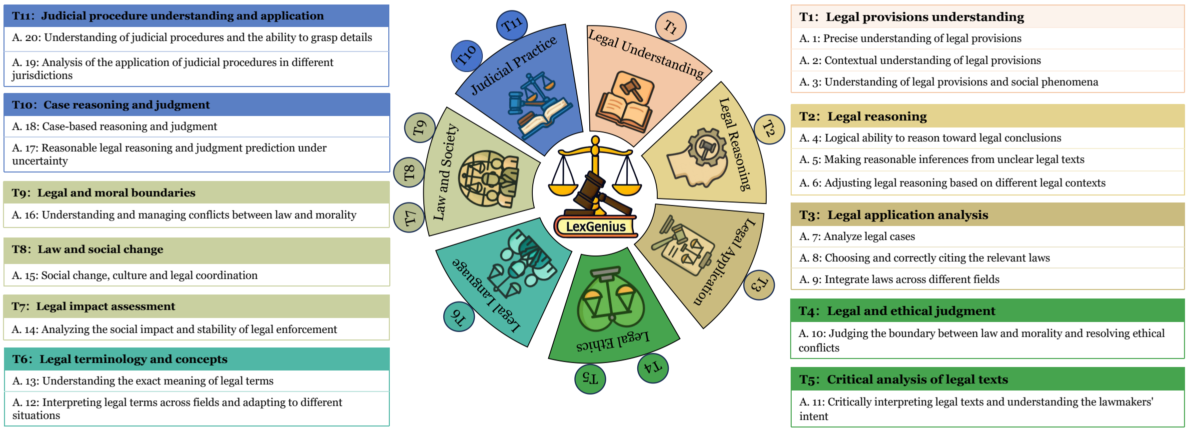


Figure 2: LexGenius can be divided into 3 levels: The first level includes **Dimensions 1-7**, the second level includes **Tasks 1-11**, and the third level includes **Abilities 1-20** (A. 1 to A. 20). Each is numbered for reference in the text.

ond half of AI, a series of expert-level benchmarks for evaluating LLMs have emerged across various domains (Cao et al., 2025; Ni et al., 2025; Li et al., 2025c,d): PhysBench (Chow et al., 2025) and PhysReason (Zhang et al., 2025c) enhance LLMs' understanding of physics; MedXpertQA (Zuo et al., 2025) and Medagentsbench (Tang et al., 2025) focus on medical knowledge; UGMathBench (Xu et al., 2025b) assesses math reasoning; and UniToMBench (Thiyagarajan et al., 2025) improves theory of mind. Benchmarks like ShotBench (Liu et al., 2025a), FinTMMBench (Zhu et al., 2025), and Chengyu-Bench (Fu et al., 2025) evaluate other fields. However, an expert-level benchmark for legal intelligence is absent (Wang et al., 2025).

3 LexGenius Framework

In this section, we outline the LexGenius framework (including seven dimensions, eleven tasks, and twenty abilities), which is shown in Figure 2.

3.1 Dimension: Education and Career Focus

The seven legal dimensions of LexGenius are based on Bloom's Taxonomy of Educational Objectives (Bloom et al., 1956), covering the cognitive hierarchy of remembering, understanding, applying, analyzing, evaluating, and creating, alongside the modular model used in legal evaluations across countries, focusing on normative understanding, rule application, procedural operation, and value judgment (Wu and Chan, 2012; Moon, 2020; Parsons et al., 2024). In the hierarchy, remembering and understanding correspond to legal understanding, applying to legal application, analyzing to legal reasoning, evaluating to legal ethics and law and society, creating to advanced arguments, legal language to clarity, and judicial practice to procedural integrity, forming a framework aligned with cognitive principles and professional needs.

3.2 Task: Theory and Practice Synergy

Further, based on Legal Hermeneutics (Leyh, 2021) and Problem-Solving Cycle theory (Stein, 1993), we decompose LexGenius's 7 legal intelligence dimensions into 11 tasks. These tasks align with common legal practice requirements and focus on textual deconstruction, case adaptation, and procedural implementation. Legal Hermeneutics guides the understanding of provisions, critical analysis of texts, and terminology, ensuring accuracy in interpretation. Problem-Solving Cycle theory simulates legal practice, driving reasoning and application analysis for problem-solving, legal and ethical judgment, moral boundaries for value calibration, case reasoning, judicial procedure understanding for validation, and legal impact and social change review, forming a task system for legal problem-solving.

3.3 Ability: Constructivist Learning-based

Furthermore, based on Constructivist Learning Theory (Ariati et al., 2025), we extract twenty atomic legal intelligence abilities from the eleven tasks. The theory shifts from outcome assessment to capturing knowledge paths through cognitive traces, simulating real legal scenarios to ensure that the evaluation reflects true professional abilities while aligning cognitive principles with occupational needs. The hierarchy from dimensions to abilities allows LexGenius to perform a detailed, multi-dimensional assessment of LLMs' legal intelligence, supporting evaluation and optimization.

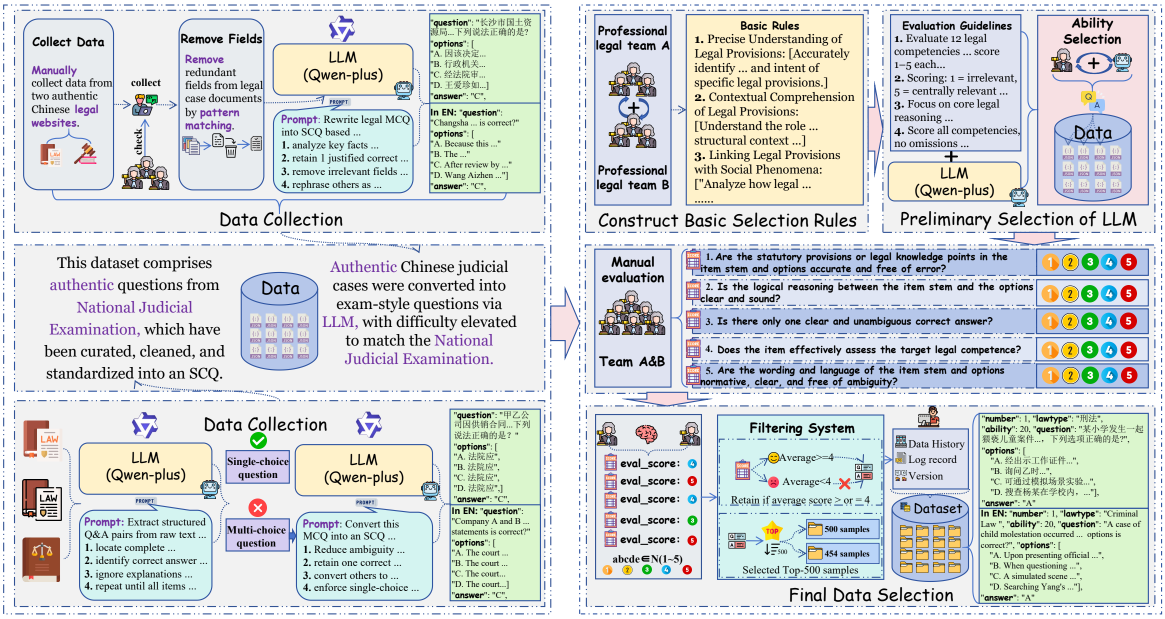


Figure 3: The MCQ construction workflow of the LexGenius, which is a process where LLM and manual work are combined. It includes three steps: data collection and structuring, construction of MCQs, and manual review.

4 LexGenius Construction

In this section, we introduce the construction principles, construction workflow, data statistics, and evaluation method of the proposed LexGenius.

4.1 Construction Principles

To avoid data leakage and contamination, LexGenius was built from scratch, using recent Chinese legal exam questions and judgment cases. We avoided reusing existing legal datasets to minimize contamination risks and ensure originality. A structured legal competency framework, developed by experienced legal experts, was used to comprehensively cover core legal intelligence abilities (Li et al., 2025f). To ensure benchmark effectiveness, a structured review process was established (Mohammadi et al., 2025; Li et al., 2025b). Reviewers are master's candidates in law, systematically trained and thoroughly familiar with key regulations, case frameworks, and legal reasoning methods.

4.2 Construction Workflow

As shown in Figure 3, the process of the construction workflow for legal QA includes three steps:

Step 1: Data collection and structuring. To ensure that the legal basis for the questions is authentic, authoritative, and semantically complete, we systematically collected the latest legal examination question banks and recent judicial cases and used LLM to clean and process these texts in a standardized manner, including encoding format conversion, removal of redundant punctuation, and paragraph reconstruction, to build a structured legal question bank and corpus. Each text is attached with a unique document identifier, source, and usage rights as metadata and saved in a unified JSON structure to facilitate index calls and traceability management when constructing legal MCQs later.

Step 2: Construction of MCQs. There are two methods for constructing multiple-choice questions, both based on the large language model. One is to screen and modify legal examination questions. Questions meeting the legal abilities are retained; others are modified using LLM. For questions related to legal soft intelligence, an LLM is used to generate them. The first method selects and modifies questions from the legal question bank: MCQs with a single correct answer are retained. In contrast, those with multiple answers are adapted using LLMs and prompt templates to ensure fairness and difficulty. For LLM-generated MCQs, we designed prompt templates with task constraints, ability descriptions, and examples to guide question generation based on specific legal cases, ensuring unique answers and a clear legal basis.

Step 3: Manual Review. To ensure legal accuracy and competency alignment, we established a team of 9 master candidates in law and created a review process. Each question undergoes double-blind scoring by 2 independent reviewers. The re-

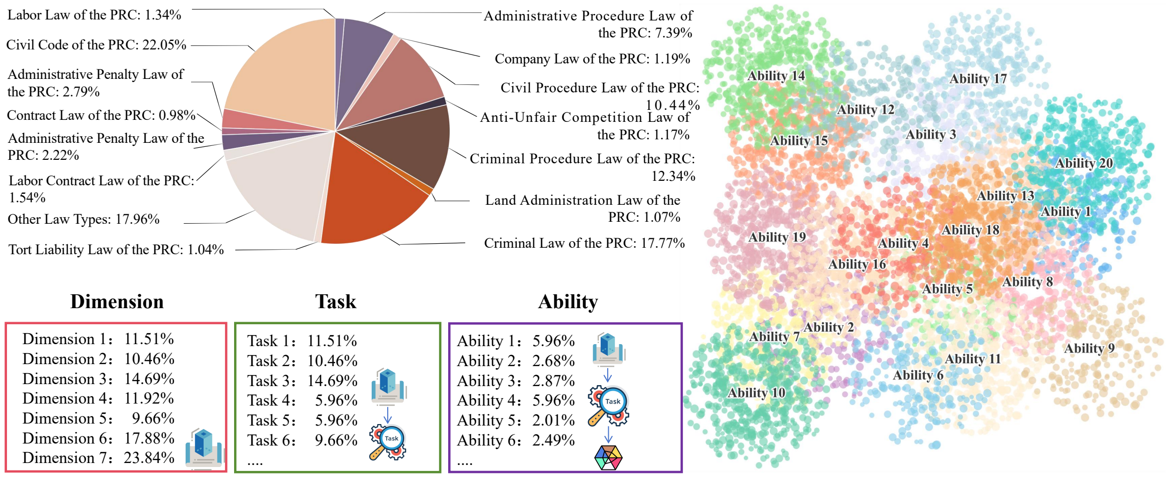


Figure 4: Data distribution of LexGenius. Left: the MCQ proportions across different laws and the dimensions, tasks, and abilities. Right: the MCQ proportions of abilities. The PRC: the People's Republic of China.

view dimensions include legal accuracy, reasoning rigor, answer uniqueness, competency alignment, and expression standardization, using a five-point scale. When there is a significant discrepancy (e.g., a difference of more than two points), a third reviewer is brought in for arbitration. Questions are retained only if their average score across all dimensions is no less than 4 points and their total score is in the top 500. The final agreement is 99.2%.

4.3 Data Statistics

After multiple rounds of review, the final version of LexGenius consists of 8,385 high-quality legal MCQs. Each question is stored in a structured JSON format, including fields such as question number, competency label, applicable law, question, options, and answers. All data versions are version-controlled, with change logs recorded during updates. To ensure explainability and accountability, we document the construction, review, and modification history of each question, supporting efficient management. Figure 4 shows the number of MCQs for each ability, covering civil disputes, corporate transactions, administrative litigation, criminal litigation, and constitutional rights.

4.4 Evaluation Method

To evaluate the LLM's legal intelligence, LexGenius adopts a three-level framework (Kahan, 2015; Huang et al., 2024). The dimension level categorizes tasks into legal cognition areas, providing insight into performance. The task level breaks dimensions into real-world tasks, testing the application, while the ability level evaluates legal abilities, identifying performance differences. At the ability level, scores $A_{i,j,k}$ are the average correctness of MCQs in each ability, where $A_{i,j,k} = \frac{1}{n_{i,j,k}} \sum_{m=1}^{n_{i,j,k}} C_{i,j,k,m}$, with $C_{i,j,k,m}$ the correctness of the m-th MCQ in the k-th ability of the j-th task in the i-th dimension, and $n_{i,j,k}$ the number of MCQs for that ability. At the task level, the task score $T_{i,j}$ is the average of ability scores within the task, calculated as $T_{i,j} = \frac{1}{m_{i,j}} \sum_{k=1}^{m_{i,j}} A_{i,j,k}$, where $m_{i,j}$ is the number of abilities in the j-th task of the i-th dimension. At the dimension level, the dimension score D_i is the average of task scores, expressed as $D_i = \frac{1}{n_i} \sum_{j=1}^{n_i} T_{i,j}$, where n_i is the number of tasks in the i-th dimension.

5 Experiments and Results

In this section, we analyze the experimental results and answer these research questions (RQs): **RQ1:** Can LLMs' legal general intelligence rival human legal experts? **RQ2:** How mature is LLMs' legal soft intelligence? **RQ3:** Do LLMs truly understand legal language? **RQ4:** Can the enhanced methods of LLMs improve their legal intelligence?

5.1 Experimental Setup

We evaluated twelve SOTA LLMs with LexGenius, which include DeepSeek-LLM-7B-Chat (DeepSeek-7B) (Bi et al., 2024), Qwen-2.5-7B-Instruct (Qwen-2.5-7B) (Hui et al., 2024), Qwen-2.5-1.5B-Instruct (Qwen-2.5-1.5B) (Hui et al., 2024), Qwen-3-8B (Yang et al., 2025), Qwen-3-4B (Yang et al., 2025), GLM-4-9B-Chat (GLM-4-9B) (GLM et al., 2024), LLaMA-3.2-1B-Instruct (LLaMA-3.2-1B) (Grattafiori et al., 2024), LLaMA-3.2-8B-Instruct (LLaMA-3.2-

Model	Legal Und.		Legal Rea.		Legal App.		Legal Ethics		Legal Lan.		Law & Soc.		Judicial Pra.		Avg.	
	Nai.	CoT	Nai.	CoT	Nai.	CoT	Nai.	CoT	Nai.	CoT	Nai.	CoT	Nai.	CoT	Nai.	CoT
Human	**89.14**		**89.14**		**89.14**		**85.19**		**87.78**		**87.41**		**86.67**		**87.78**	
Qwen2.5-1.5B	44.05	43.93	41.01	42.28	35.61	35.83	55.70	56.00	51.98	53.32	59.33	59.80	45.50	46.35	47.60	48.22
Qwen2.5-7B	61.41	60.50	57.42	56.57	45.26	45.19	64.00	63.60	66.08	65.92	66.07	66.27	57.55	57.60	59.68	59.38
Qwen3-4B	53.23	52.43	50.37	50.57	39.05	38.07	60.90	61.00	57.63	57.36	62.20	61.13	51.55	51.75	53.56	53.19
Qwen3-8B	58.19	58.77	52.07	52.72	38.73	37.21	62.90	61.60	63.09	61.93	61.20	60.47	55.35	54.40	55.93	55.30
LLaMA-3.2-1B	29.37	28.33	24.03	25.84	26.47	26.23	44.40	43.40	39.37	39.93	50.27	47.93	33.05	32.25	35.28	34.85
LLaMA-3.2-8B	36.99	35.66	31.38	32.85	33.20	33.40	53.50	53.10	46.38	46.62	55.40	56.47	42.80	43.55	42.81	43.09
GLM-4-9B	52.85	53.05	47.78	47.54	35.52	36.47	61.40	61.40	62.46	63.24	61.87	61.47	50.55	51.05	53.20	53.46
DeepSeek-7B	37.27	33.56	34.28	30.96	29.54	30.88	45.10	43.30	39.53	40.79	47.00	47.40	37.20	36.55	38.56	37.63
DeepSeek-R1	**67.35**	**67.76**	**61.18**	<u>61.50</u>	<u>45.91</u>	<u>45.90</u>	**70.80**	**70.60**	**71.10**	**71.20**	**72.40**	<u>72.33</u>	**62.40**	**62.85**	**64.45**	**64.59**
DeepSeek-V3	<u>67.04</u>	<u>67.55</u>	<u>60.90</u>	**61.57**	**46.36**	**46.60**	<u>70.30</u>	<u>69.40</u>	<u>70.48</u>	<u>70.52</u>	<u>71.73</u>	**72.60**	**62.40**	<u>62.50</u>	<u>64.17</u>	<u>64.39</u>
GPT-4o mini	49.77	49.80	43.26	43.91	36.99	36.44	62.20	62.20	57.51	57.53	65.47	65.80	52.25	52.60	52.49	52.61
GPT-4.1 nano	46.51	54.62	43.23	43.88	30.92	40.05	56.80	59.60	53.11	62.69	58.80	60.53	48.20	52.60	48.22	53.43

Table 2: Comparison of Naive (Nai.) and CoT prompts of LLMs on LexGenius (all values in %). Bold entries are the best results with the Naive (CoT) prompt; Underlined entries are the 2nd-best with the Naive (CoT) prompt. (Legal Und. means Legal Understanding; Legal Rea. means Legal Reasoning; Legal App. means Legal Application; Legal Lan. means Legal Language; Law & Soc. means Law and Society; and Judicial Pra. means Judicial Practice.)

8B) (Grattafiori et al., 2024), DeepSeek-R1 (Guo et al., 2025), DeepSeek-V3 (Liu et al., 2024), GPT-4o mini (Hurst et al., 2024), and GPT-4.1 nano (Hurst et al., 2024). We followed the official protocols, using official APIs or LLM weights where applicable. The evaluation utilized two types of prompts: the first type was the Naive prompt; the second was the CoT prompt, encouraging the LLMs to perform step-by-step reasoning. To prevent potential bias, we shuffled the answer options twice and averaged the scores of each LLM.

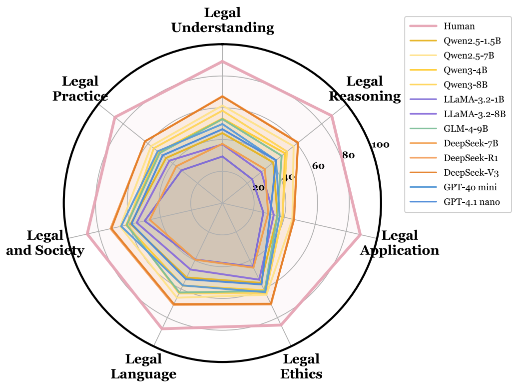


Figure 5: Comparison of the 12 SOTA LLMs with human experts on 7 core dimensions of legal intelligence.

5.2 Main Results (RQ1)

The performance of the twelve SOTA LLMs across seven dimensions (see Table 2 and Figure 5), eleven tasks (see Table 3), and twenty ability rankings (see Figure 6) on LexGenius is reported.

Comparison with Human. As shown in Table 2 and Figure 5, although LLMs excel in generating legal texts, their capabilities across the seven dimensions still fall short compared to human experts, particularly in areas like legal reasoning, judicial practice, and legal ethics, where value judgments and contextual trade-offs are key. This underscores that legal intelligence is not just about reciting rules but about making sound judgments amidst uncertainty, relying on human experiences, ethical intuition, and institutional understanding. While LLMs can articulate legal principles, they are not yet capable of rendering nuanced judgments. They are powerful assistants, not true counterparts.

Task-view performance. As shown in Table 3, LLMs perform relatively well in static knowledge-based tasks (e.g., legal provisions understanding). However, they are significantly weaker than human experts in tasks that require dynamic reasoning and institutional understanding (e.g., legal application analysis and case reasoning and judgment). Particularly in tasks involving value trade-offs (e.g., legal and ethical judgment), LLMs tend to avoid complex judgments and lack critical thinking and contextual sensitivity. This indicates that they still lack the comprehensive judgment capabilities required for legal practice and remain tools for assistance rather than equivalent intelligent agents.

Ranking of LLMs. Figure 6 reveals that the LLMs' average scores and rankings are nearly identical. Only a few leading models perform comprehensively and stably, while most rank lower with similar capabilities. This head convergence and tail dispersion pattern suggests that current large models lack balanced legal general intelligence. Their strengths lie in formalized tasks, but they remain weak in complex abilities requiring cross-dimensional integration, value judgment, or institu-

Model	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Avg.
Human	**89.14**	**89.14**	**89.14**	**86.67**	**83.70**	**87.78**	**87.41**	**91.85**	**82.96**	**90.00**	**83.34**	**87.37**
Qwen2.5-1.5B	44.05	41.01	35.61	59.00	52.40	51.98	57.20	56.40	64.40	44.60	46.40	50.28
Qwen2.5-7B	61.41	57.42	45.26	64.80	63.20	66.08	71.00	58.00	69.20	55.20	59.90	61.04
Qwen3-4B	53.23	50.37	39.05	58.00	63.80	57.63	69.00	55.60	62.00	49.50	53.60	55.62
Qwen3-8B	58.19	52.07	38.73	64.20	61.60	63.09	67.60	54.80	61.20	55.50	55.20	57.47
LLaMA-3.2-1B	29.37	24.03	26.47	46.80	42.00	39.37	45.80	46.00	59.00	33.90	32.20	38.63
LLaMA-3.2-8B	36.99	31.38	33.20	56.20	50.80	46.38	55.00	51.60	59.60	43.90	41.70	46.07
GLM-4-9B	52.85	47.78	35.52	66.00	56.80	62.46	64.40	57.60	63.60	49.70	51.40	55.28
DeepSeek-7B	37.27	34.28	29.54	48.40	41.80	39.53	51.60	40.20	49.20	36.90	37.50	40.57
DeepSeek-R1	**67.35**	**61.18**	45.91	**67.60**	**74.00**	**71.10**	**76.80**	65.40	**75.00**	**60.60**	**64.20**	**66.29**
DeepSeek-V3	67.04	60.90	**46.36**	**67.60**	73.00	70.48	76.40	**66.20**	72.60	**60.60**	**64.20**	65.94
GPT-4o mini	49.77	43.26	36.99	65.20	59.20	57.51	67.60	61.20	67.60	50.30	54.20	55.71
GPT-4.1 nano	46.51	43.23	30.92	59.20	54.40	53.11	63.00	54.40	59.00	48.00	48.40	50.92

Table 3: Performance of twelve LLMs and human experts on eleven legal tasks, showing a significant gap between LLMs and humans. DeepSeek-R1 and DeepSeek-V3 are the top performers, with the greatest challenge in Task 3.

Model	A.10	A.11	A.13	A.14	A.15	A.16	Avg.
Human	**86.7**	**83.7**	**85.9**	**87.4**	**91.9**	**83.0**	**86.4**
Qwen2.5-1.5B	59.0	52.4	60.4	57.2	56.4	64.4	58.3
Qwen2.5-7B	64.8	63.2	68.6	71.0	58.0	69.2	65.8
Qwen3-4B	58.0	63.8	64.6	69.0	55.6	62.0	62.2
Qwen3-8B	64.2	61.6	65.2	67.6	54.8	61.2	62.4
LLaMA-3.2-1B	46.8	42.0	51.0	45.8	46.0	59.0	48.4
LLaMA-3.2-8B	56.2	50.8	63.4	55.0	51.6	59.6	56.1
GLM-4-9B	66.0	56.8	66.2	64.4	57.6	63.6	62.4
DeepSeek-7B	48.4	41.8	46.8	51.6	40.2	49.2	46.3
DeepSeek-R1	**67.6**	**74.0**	68.0	**76.8**	65.4	**75.0**	**71.1**
DeepSeek-V3	**67.6**	73.0	67.4	76.4	**66.2**	72.6	70.5
GPT-4o mini	65.2	59.2	**69.2**	67.6	61.2	67.6	65.0
GPT-4.1 nano	59.2	54.4	60.4	63.0	54.4	59.0	58.4

Table 4: Comparison of the twelve SOTA LLMs for legal soft intelligence on LexGenius. (A. means Ability.)

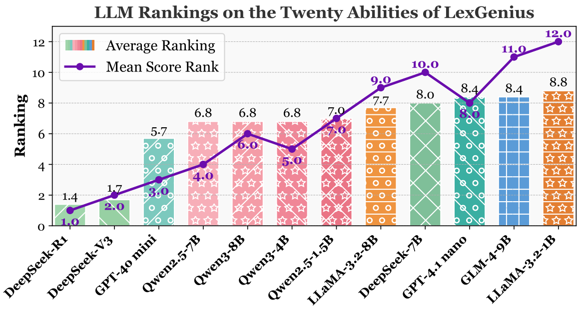


Figure 6: Average ranking and average score ranking of the 12 SOTA LLMs in the 20 legal intelligence abilities.

tional understanding. Even top models approaching human-level performance still lack the deep coupling and contextual adaptability required for legal practice, falling short of experts' capabilities.

Case study. This case (see Figure 8) highlights LLMs' limitations in legal reasoning: their judgments rely on surface cues, oversimplifying rights conflicts and ignoring core context. While DeepSeek-R1 anchors personality rights, GPT-4o mini misjudges liability, showing a lack of holistic legal understanding. LLMs remain trapped in decontextualized reasoning, unable to balance norms, facts, and values like experts. The gap lies in contextual balancing, a key aspect of legal intelligence.

Naive prompt vs CoT prompt. While CoT enhances surface-level reasoning in several LLMs (see Table 2), it exposes their limitations in high-level legal intelligence (e.g., application, ethics, and judicial practice). The improvement stops at formal logic, failing to capture the nuanced judgments made by human experts that integrate norms, context, and ethics. Humans navigate complexity with stability, while models remain confined to static knowledge reorganization. This highlights that the gap in legal intelligence lies not in reasoning but in making responsible decisions amidst uncertainty, an area current LLMs have yet to bridge.

5.3 Legal Soft Intelligence Analysis (RQ2)

As shown in Table 4, results reveal systematic immaturity in LLMs' legal soft intelligence: LLMs show significant gaps in higher-order abilities like social change, culture, legal coordination, and law-morality boundaries, with fewer issues in analyzing legal enforcement's social impact. This reflects deficits in experiential social knowledge and ethical reasoning. Scaling fails to overcome performance ceilings, revealing architectural limits in acquiring moral intuition and judgment from static text.

5.4 Legal Language Mastery Analysis (RQ3)

We evaluated the performance of LLMs across nine legal language abilities. The results (see Figure 7) show LLMs excel at reproducing legal text structure and procedural patterns, performing well on formatted tasks. However, their abilities degrade

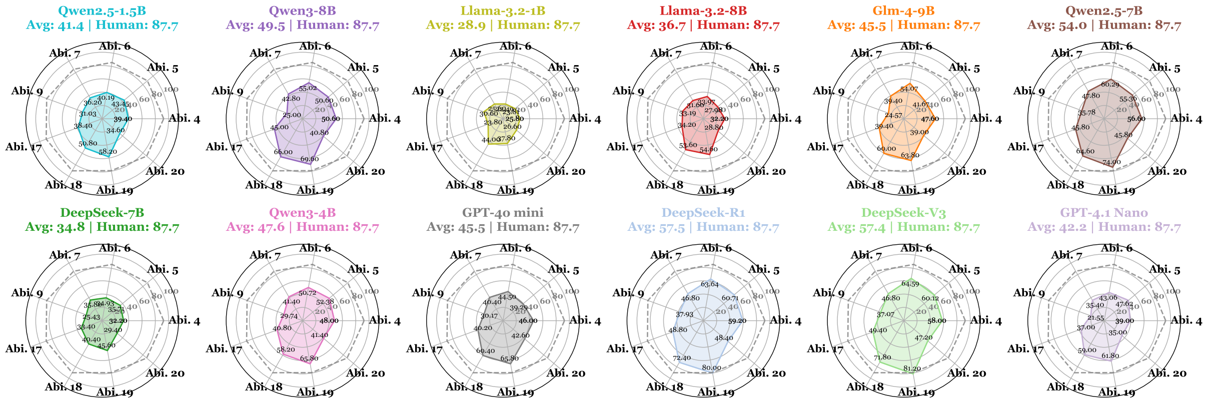


Figure 7: Performance of 12 LLMs across 6 legal language indicators, showing gaps compared to the human baseline. Even the best-performing LLMs (e.g., Deepseek-R1) fall short in mastering legal language. Abi. is Ability.

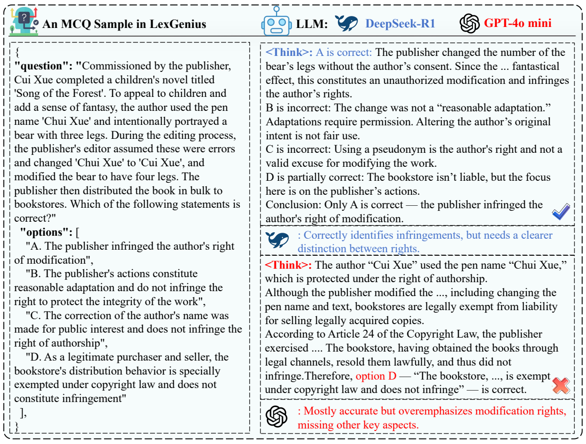


Figure 8: We utilize an MCQ sample case to evaluate DeepSeek-R1 and GPT-4o mini and present the respective thought processes of both LLMs. The English translation of the original Chinese test sample is on the left.

significantly when faced with ambiguity, conflict, or value trade-offs in real legal reasoning. This gap arises from an inherent limitation: models lack understanding of institutional logic, social context, and ethical goals, relying solely on statistical correlations. As a result, LLMs replicate the form of law without truly understanding it and can assist but not replace the essential normative insight and value judgments in legal decision-making.

5.5 With Different Enhanced Methods (RQ4)

As shown in Table 5, the comparison results (more details see Appendix F) reveal a triple decoupling phenomenon in LLM legal intelligence: *(i)* Scale-performance decoupling shows that legal intelligence exhibits a non-monotonic relationship with model size and family, suggesting that scale alone is insufficient for legal-task performance. *(ii)* Reasoning paradigm decoupling shows that CoT fails to consistently improve over baseline and often underperforms it, indicating a mismatch with the deterministic and constrained nature of legal tasks. *(iii)* Optimization strategy decoupling shows that SFT, RAG, and GRPO yield divergent outcomes across models: SFT and GRPO both improve the other three models but underperform the strongest baseline model (Qwen2.5-7B), while RAG improves only Qwen3 models and harms Qwen2.5 models, suggesting that knowledge retrieval and legal reasoning capacity are not complementary.

Model	Baseline	CoT	SFT	RAG	GRPO
Qwen2.5-1.5B	48.53	49.44	51.49	34.66	52.48
Qwen2.5-7B	58.45	57.16	55.86	52.22	55.93
Qwen3-4B	29.32	28.39	51.30	37.67	51.01
Qwen3-8B	27.86	27.79	56.84	45.78	55.04

Table 5: Comparison of four LLMs on LexGenius with the enhanced methods, including CoT, Supervised Fine-Tuning (SFT), Retrieval-Augmented Generation (RAG), and Group Relative Policy Optimization (GRPO).

6 Conclusion

In this work, we propose LexGenius, an expert-level and comprehensive benchmark for evaluating LLMs' legal general intelligence capabilities. Based on the three-level framework (Dimension–Task–Capability), we assess twelve SOTA LLMs from different perspectives. LexGenius addresses gaps in existing benchmarks, including systematic evaluation and alignment with real-world legal reasoning. Experimental results reveal significant legal intelligence gaps of LLMs, highlighting disparities with human legal experts and their specific weaknesses in legal general intelligence.

Limitations

In Appendix H, we discuss the limitations of LexGenius. Furthermore, we outline the future work of LexGenius in Appendix I.

Ethical Considerations

This work complies with the ACL Ethics Policy, relying on anonymized, publicly available legal resources to ensure privacy and academic integrity.

Acknowledgments

This work was supported by the National Key Research and Development Program of China [grant number 2023YFC3304903]

References

Jati Ariati, Thomas Pham, and Jane S Vogler. 2025. Constructivist learning environments: Validating the community of inquiry survey for face-to-face contexts. *Active Learning in Higher Education*, 26(1):41–57.

Arian Askari, Suzan Verberne, and Gabriella Pasi. 2022. Expert finding in legal community question answering. In *European conference on information retrieval*, pages 22–30. Springer.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Benjamin S Bloom, MD Engelhart, EJ Furst, WH Hill, and DR Krathwohl. 1956. Taxonomy of educational objectives: Theclassification of educational goals. *Handbook I: Cognitive domain. New York: David McKay Company*.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. 2024. XAI meets LLMs: A survey of the relation between explainable AI and large language models. *arXiv preprint arXiv:2407.15248*.

Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, et al. 2025. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838*.

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

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

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

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. 2025. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*.

Maelenn Corfmat, Joé T Martineau, and Catherine Régis. 2025. High-reward, high-risk technologies? an ethical and legal account of ai development in healthcare. *BMC medical ethics*, 26(1):4.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2023. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv preprint arXiv:2306.16092*.

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 *COLING*.

Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, et al. 2025. Safeguarding large language models: A survey. *Artificial Intelligence Review*, 58(12):382.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, et al. 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.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. Chengyu-bench: Benchmarking large language models for chinese idiom understanding and use. *arXiv preprint arXiv:2506.18105*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

George Hannah, Rita T Sousa, Ioannis Dasoulas, and Claudia d'Amato. 2025. On the legal implications of large language model answers: A prompt engineering approach and a view beyond by exploiting knowledge graphs. *Journal of Web Semantics*, 84:100843.

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

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

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zheng Jia, Shengbin Yue, Wei Chen, Siyuan Wang, Yidong Liu, Yun Song, and Zhongyu Wei. 2025. Ready jurist one: Benchmarking language agents for legal intelligence in dynamic environments. *arXiv preprint arXiv:2507.04037*.

Dan M Kahan. 2015. Laws of cognition and the cognition of law. *Cognition*, 135:56–60.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51(3):371–402.

Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharrer, Megan Ma, and Marzieh Nabi. 2025. Towards robust legal reasoning: Harnessing logical llms in law. *arXiv preprint arXiv:2502.17638*.

Junghwan Kim, Hyeonseok Jeon, Dongseok Heo, Jung Lee, and Bongwon Suh. 2025. Legisflow: Enhancing korean legal research with temporal-aware llm interfaces. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–29.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Gregory Leyh. 2021. *Legal hermeneutics: history, theory, and practice*. University of California Press.

Ang Li, Yiquan Wu, Yifei Liu, Ming Cai, Lizhi Qing, Shihang Wang, Yangyang Kang, Chengyuan Liu, Fei Wu, and Kun Kuang. 2025a. Unilr: Unleashing the power of llms on multiple legal tasks with a unified legal retriever. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11953–11967.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2025b. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. 2025c. 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, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024a. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *arXiv preprint arXiv:2409.20288*.

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.

Jiawei Li, Yang Gao, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, et al. 2025d. Fundamental capabilities and applications of large language models: A survey. *ACM Computing Surveys*.

Shangyuan Li, Shiman Zhao, Zhuoran Zhang, Zihao Fang, Wei Chen, and Tengjiao Wang. 2025e. Basis is also explanation: Interpretable legal judgment reasoning prompted by multi-source knowledge. *Information Processing & Management*, 62(3):103996.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025f. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1587–1606.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Yangguang Li, Weichao Chen, et al. 2025a. Shotbench: Expert-level cinematic understanding in vision-language models. *arXiv preprint arXiv:2506.21356*.

Qian Liu, Hang Yu, Qiqi Wang, Qi Xu, Jinpeng Li, Zhuoqun Zou, Rui Mao, and Erik Cambria. 2025b. Legal knowledge infusion for large language models: A survey. *Information Fusion*, page 103426.

Wenjin Liu, Haoran Luo, Xueyuan Lin, Haoming Liu, Tiesunlong Shen, Jiapu Wang, Rui Mao, and Erik Cambria. 2025c. Prompt-r1: Collaborative automatic prompting framework via end-to-end reinforcement learning. *arXiv preprint arXiv:2511.01016*.

Haoran Luo, Guanting Chen, Qika Lin, Yikai Guo, Fangzhi Xu, Zemin Kuang, Meina Song, Xiaobao Wu, Yifan Zhu, Luu Anh Tuan, et al. 2025a. Graph-r1: Towards agentic graphrag framework via end-to-end reinforcement learning. *arXiv preprint arXiv:2507.21892*.

Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Luu Anh Tuan. 2025b. Hypergraphrag: Retrieval-augmented generation via hypergraph-structured knowledge representation.

Haoran Luo, Haihong E, Yikai Guo, Qika Lin, Xiaobao Wu, Xinyu Mu, Wenhao Liu, Meina Song, Yifan Zhu, and Luu Anh Tuan. 2025c. Kbqa-o1: Agentic knowledge base question answering with monte carlo tree search.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. GPTEval: A survey on assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and benchmarking of llm agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6129–6139.

Caitlin Moon. 2020. The case for the delta model for lawyer competency. *Law Prac.*, 46:38.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, et al. 2025. A survey on large language model benchmarks. *arXiv preprint arXiv:2508.15361*.

Patrick Parsons, Michelle Hook Dewey, and Kristina L Niedringhaus. 2024. Georgia state legal technology competency model: A framework for examining and evaluating what it means to be a technologically competent lawyer. *U. St. Thomas LJ*, 20:53.

Tiesunlong Shen, Erik Cambria, Jin Wang, Yi Cai, and Xuejie Zhang. 2025. Insight at the right spot: Provide decisive subgraph information to graph llm with reinforcement learning. *Information Fusion*, 117:102860.

Barry S Stein. 1993. *The IDEAL problem solver: A guide for improving thinking, learning, and creativity*. WH Freeman.

Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Zibing Que, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. Stard: A chinese statute retrieval dataset with real queries issued by non-professionals. *arXiv preprint arXiv:2406.15313*.

Weihang Su, Baoqing Yue, Qingyao Ai, Yiran Hu, Jiaqi Li, Changyue Wang, Kaiyuan Zhang, Yueyue Wu, and Yiqun Liu. 2025. Judge: Benchmarking judgment document generation for chinese legal system. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3573–3583.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. 2025. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Prameshwar Thiyagarajan, Vaishnavi Parimi, Shamant Sai, Soumil Garg, Zhangir Meirbek, Nitin Yarlagadda, Kevin Zhu, and Chris Kim. 2025. Unitombench: Integrating perspective-taking to improve theory of mind in llms. *arXiv preprint arXiv:2506.09450*.

Jiaqi Wang, Huan Zhao, Zhenyuan Yang, Peng Shu, Junhao Chen, Haobo Sun, Ruixi Liang, Shixin Li, Pengcheng Shi, Longjun Ma, et al. 2024. Legal evaluations and challenges of large language models. *arXiv preprint arXiv:2411.10137*.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.

Richard WS Wu and Kay-Wah Chan. 2012. Regulatory regimes for lawyers' ethics in japan and china: A comparative study. *Tsinghua China L. Rev.*, 5:49.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. 2025. AntiLeak-Bench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Vienna, Austria. Association for Computational Linguistics.

Huiyuan Xie, Chenyang Li, Huining Zhu, Chubin Zhang, Yuxiao Ye, Zhenghao Liu, and Zhiyuan Liu. 2025. Lawchain: Modeling legal reasoning chains for chinese tort case analysis. *arXiv preprint arXiv:2510.17602*.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025a. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Xin Xu, Jiaxin Zhang, Tianhao Chen, Zitong Chao, Jishan Hu, and Can Yang. 2025b. Ugmathbench: A diverse and dynamic benchmark for undergraduate-level mathematical reasoning with large language models. *arXiv preprint arXiv:2501.13766*.

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

Rujing Yao, Yang Wu, Chenghao Wang, Jingwei Xiong, Fang Wang, and Xiaozhong Liu. 2025a. Elevating legal llm responses: Harnessing trainable logical structures and semantic knowledge with legal reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5630–5642.

Rujing Yao, Yiquan Wu, Tong Zhang, Xuhui Zhang, Yuting Huang, Yang Wu, Jiayin Yang, Changlong Sun, Fang Wang, and Xiaozhong Liu. 2025b. Intelligent legal assistant: An interactive clarification system for legal question answering. *arXiv preprint arXiv:2502.07904*.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, and Yanqing An. 2024a. A circumstance-aware neural framework for explainable legal judgment prediction. *IEEE Transactions on Knowledge and Data Engineering*.

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

Jian Zhang, Zhangqi Wang, Haiping Zhu, Kangda Cheng, Kai He, Bo Li, Qika Lin, Jun Liu, and Erik Cambria. 2026a. Mars: Multi-agent adaptive reasoning with socratic guidance for automated prompt optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16307–16315.

Jian Zhang, Zhiyuan Wang, Zhangqi Wang, Fangzhi Xu, Qika Lin, Lingling Zhang, Rui Mao, Erik Cambria, and Jun Liu. 2026b. Maps: Multi-agent personality shaping for collaborative reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(19):16316–16324.

Kepu Zhang, Weijie Yu, Sunhao Dai, and Jun Xu. 2025a. Citalaw: Enhancing llm with citations in legal domain. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11183–11196.

Kepu Zhang, Weijie Yu, Zhongxiang Sun, and Jun Xu. 2025b. Syler: A framework for explicit syllogistic legal reasoning in large language models. *arXiv preprint arXiv:2504.04042*.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. 2025c. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*.

Yue Zhang, Zhiliang Tian, Shicheng Zhou, Haiyang Wang, Wenqing Hou, Yuying Liu, Xuechen Zhao, Minlie Huang, Ye Wang, and Bin Zhou. 2025d. Rljp: Legal judgment prediction via first-order logic rule-enhanced with large language models. *arXiv preprint arXiv:2505.21281*.

Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. 2025. Towards lifelong learning of large language models: A survey. *ACM Computing Surveys*, 57(8):1–35.

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

Zhi Zhou, Kun-Yang Yu, Shi-Yu Tian, Xiao-Wen Yang, Jiang-Xin Shi, Pengxiao Song, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2025. Lawgpt: Knowledge-guided data generation and its application to legal llm. *arXiv preprint arXiv:2502.06572*.

Fengbin Zhu, Junfeng Li, Liangming Pan, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. 2025. Fintmmbench: Benchmarking temporal-aware multi-modal rag in finance. *arXiv preprint arXiv:2503.05185*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

Appendix

A Motivation and Theoretical basis of LexGenius

In this section, we primarily explain the design motivations of the LexGenius.

A.1 Design motivation

In this section, we mainly introduce the motivation and reasons for designing LexGenius and answer the question: **Why is a structured legal general intelligence evaluation framework needed?**

Legal general intelligence is not a stack of tasks, but a simulation of cognitive collaborative chains. Existing frameworks often focus on classification or question-answering tasks, presenting only macro-level accuracy on isolated benchmarks. Such metrics fail to pinpoint errors within the complex cognitive chain of legal decision-making. Legal intelligence is not a sum of discrete tasks but an organic coordination of systematic capabilities, spanning statutory interpretation, fact extraction, rule adaptation, ethical judgment, and social impact assessment. LLM performance should therefore be decomposed into a multi-stage flow, rather than treated as monolithic. Accordingly, effective evaluation must map this cognitive chain, avoiding the pitfall of reducing legal intelligence to task-solving while ignoring how the model thinks.

From performance reporting to capability diagnosis and explanation. Traditional metrics like accuracy or F1 indicate output correctness, but they fail to address a fundamental question: At which cognitive stage did the model fail? Was it a semantic misunderstanding? Rule misapplication? Or an ethical blind spot? LexGenius introduces 20 atomic legal general intelligence abilities and establishes an interpretable, traceable, and auditable evaluation system through a tri-level mapping mechanism across the ability layer, task layer, and cognitive dimension. This structure reveals deficiencies in specific micro-level abilities and provides an actionable feedback loop for capability-oriented training, prompt optimization, and safety enhancement.

Enabling cross-stage cognitive analysis and transferability research. The complexity of legal reasoning lies in its chained structure: statutory semantics → case fact abstraction → rule application → ethical judgment → precedent alignment. Frameworks failing to distinguish performance across these stages cannot support research in multi-hop reasoning, chain-of-thought attention, or multi-task learning. Through hierarchical abstraction and modular decomposition, our framework standardizes this cognitive pathway, offering a clear baseline for investigating knowledge transfer and generalizable reasoning capabilities.

Toward a professional-grade legal general intelligence evaluation paradigm. As LLMs approach the professional thresholds of bar examinations and real-world legal practice, evaluation frameworks must likewise advance to a professional-grade level. LexGenius draws on standards from the National Legal Professional Qualification Examination and international bar exams, distinguishing specialized dimensions of legal competence, such as legal semantic understanding, norm alignment, ethical judgment, and procedural compliance. This framework breaks from the general-purpose perspective of traditional NLP benchmarks, aligning with the practical demands of legal work. Such a professional, ability-oriented, structured evaluation reveals current model boundaries and illuminates potential development trajectories.

A.2 Framework hierarchy and implementation

The structural design of the LexGenius is illustrated in Table 6. This framework supports both vertical capability dissection—capturing a model's progressive performance across stages such as legal language understanding → case application → judgment prediction—and horizontal comparison, such as evaluating differences between LLM A and LLM B along the dimension of ethical judgment.

A.3 Toward Legal Cognitive Modeling

Across various domains, an increasing number of expert-level benchmarks are emerging to advance the development and understanding of LLMs. In the legal field, the need for a benchmark that rigorously evaluates expert-level legal general intelligence is equally critical. The proposed three-tier structure—Dimension–Task–Ability—functions not only as an evaluation framework but also as a cognitive modeling paradigm. We move beyond merely assessing outcomes to examining whether a model can think, interpret, and judge like a trained legal professional when confronted with legal contexts. In this sense, LexGenius is not just a benchmark—it is a foundation designed to drive the evolution of legal general intelligence.

Level	Description	Goals and Functions	Number
Dimension	Core legal intelligence focus areas and judgments (e.g., legal reasoning)	Build top-level cognition and capability aggregation	7
Task	Scenario-based legal tasks (e.g., case reasoning and judgment prediction)	Mid-level structure linking abilities and test design	11
Ability	Measurable legal intelligence ability units (e.g., statute understanding and interpretation)	Minimum unit, supporting fine-grained evaluation and diagnosis	20

Table 6: Hierarchical levels of the LexGenius and corresponding implementation counts. It includes the Dimension level (high-level cognitive targets), the Task level (scenario-based applications), and the Ability level (fine-grained evaluable units), along with the number of implemented benchmarks under each category.

B Definitions of Legal Intelligence Abilities

This section provides detailed definitions for the twenty atomic legal intelligence abilities in the proposed LexGenius framework, following the ability names utilized in Figure 2. Each ability represents a measurable unit of legal general intelligence, assessed through standardized multiple-choice questions.

1. Precise understanding of legal provisions. Ability to accurately interpret key terms, conditions, and structural logic of legal clauses, including scope and applicability. An MCQ sample of this ability in the LexGenius is shown in Figure 9.

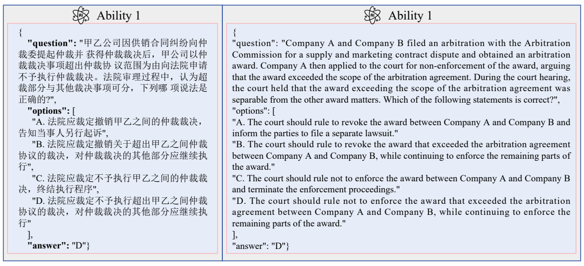


Figure 9: The MCQ sample of ability 1. The left is the original text, and the right is the English translation.

2. Contextual understanding of legal provisions. Ability to interpret legal text within the correct legal and social context, avoiding misinterpretation based on literal reading alone. An MCQ sample of this ability in the LexGenius is shown in Figure 10.

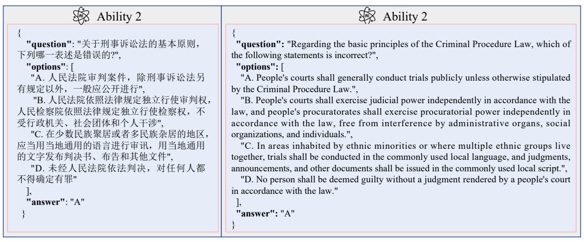


Figure 10: The MCQ sample of ability 2. The left is the original text, and the right is the English translation.

3. Understanding of legal provisions and social phenomena. Ability to relate legal provisions to real-world events, social needs, and historical developments. An MCQ sample of this ability in the LexGenius is shown in Figure 11.

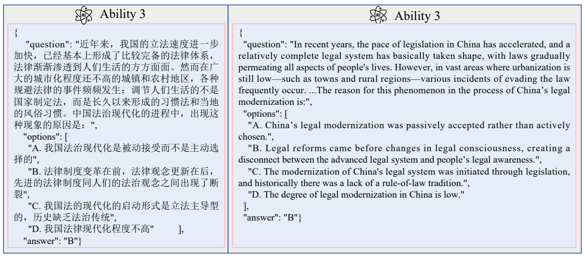


Figure 11: The MCQ sample of ability 3. The left is the original text, and the right is the English translation.

4. Logical ability to reason toward legal conclusions. Ability to construct sound legal arguments based on facts and rules, forming consistent and well-structured conclusions. An MCQ sample of this ability in the LexGenius is in Figure 12.

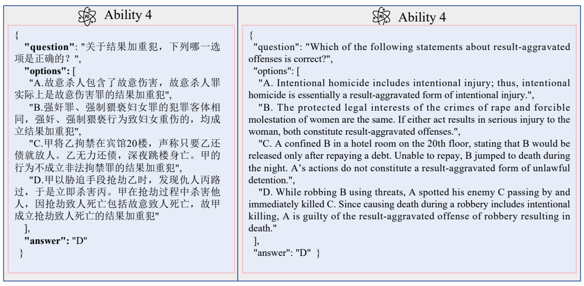


Figure 12: The MCQ sample of ability 4. The left is the original text, and the right is the English translation.

5. Making reasonable inferences from unclear legal texts. Ability to infer appropriate meanings from vague, ambiguous, or abstract legal language using legal logic and principles. An MCQ sample of this ability in the LexGenius is in Figure 13.

Ability 5


```
{
  "question": "言论自由是公民监督和制约公权力的重要手段。现代法治国家，往往以言论自由作为衡量一个国家法治程度的标杆，但是言论自由也有其界限。下列哪一行为超出了言论自由的范畴?",
  "options": [
    "A. 揭发某国家机关领导包养情妇，将掌握的证据刊登在报纸上",
    "B. 在电视购物节目中，商家为了推广自己的产品，公开制作秘方",
    "C. 军队领导在接受记者采访时，向记者透露我国核弹头储备的数量",
    "D. 某学者在报纸上公开批评他人作品，言辞激烈，极尽嘲讽"
  ],
  "answer": "C"
}
```


Ability 5


```
{
  "question": "Freedom of speech is an important means by which citizens supervise and restrain public power. In modern rule-of-law countries, freedom of speech is often regarded as a benchmark for evaluating the degree of legal development. However, freedom of speech also has its boundaries. Which of the following actions exceeds the scope of freedom of speech?",
  "options": [
    "A. Exposing a government official for keeping a mistress and publishing the evidence in a newspaper.",
    "B. A merchant publicly discloses a secret formula in a TV shopping program to promote their product.",
    "C. A military leader reveals the number of nuclear warheads China possesses during a media interview.",
    "D. A scholar publicly criticizes another's work in a newspaper with harsh and mocking language."
  ],
  "answer": "C"
}
```


Figure 13: The MCQ sample of ability 5. The left is the original text, and the right is the English translation.

6. Adjusting legal reasoning based on different legal contexts. Ability to adapt reasoning strategies when applying different branches of law, such as civil, criminal, or administrative. An MCQ sample of this ability in the LexGenius is shown in Figure 14.

Ability 6


```
{
  "question": "某商业公司有一座7层的办公楼，2022年7月， 某商业公司向H 银行贷款，并将这座办公楼抵押给 H 银行作担保。2022年12月，某商业公司因一购销合同与某贸易有限公司发生纠纷，某贸易有限公司向法院起诉，并申请保全。受诉法院裁定将某商业公司办公楼查封。下列选项哪个是正确的：",
  "options": [
    "A.法院所作的裁定是错误的，因该办公楼已事先 抵押给了H 银行",
    "B. 法院可以作出保全裁定，但事先应征得H 银行 的同意",
    "C. 法院可以作出保全裁定，但 H 银行对该办公楼 仍有优先受偿权",
    "D. 法院作出裁定是正确的，H 银行因此对该办公 楼丧失优先受偿权"
  ],
  "answer": "C"
}
```


Ability 6


```
{
  "question": "A commercial company owns a 7-story office building. In July 2022, the company obtained a loan from H Bank and mortgaged the office building as collateral. In December 2022, the company became involved in a dispute with a trading limited company over a purchase and sales contract. The trading company filed a lawsuit and applied for property preservation. The court ruled to seize the office building. Which of the following statements is correct?",
  "options": [
    "A. The court's ruling is incorrect because the office building had already been mortgaged to H Bank.",
    "B. The court may issue a property preservation ruling, but must first obtain consent from H Bank.",
    "C. The court may issue a property preservation ruling, but H Bank still retains priority claim over the office building.",
    "D. The court's ruling is correct, and H Bank thereby loses its priority claim over the office building."
  ],
  "answer": "C"
}
```


Figure 14: The MCQ sample of ability 6. The left is the original text, and the right is the English translation.

7. Analyze legal cases. Ability to identify relevant facts and legal issues in a case and link them with the applicable legal norms or precedents. An MCQ sample of this ability in the LexGenius is shown in Figure 15.

Ability 7


```
{
  "question": "外卖骑手张某见李某跳河，将手机交给路人王某便从十米高跳入湖中救李某，跳下去背部不慎划伤，李某不想被救拼命挣脱，张某因用力过大导致背部伤口撕扯，李某因此胳膊骨折，王某因不小心将张某手机掉在地上，屏幕被摔碎。关于本案，下列说法正确的是？",
  "options": [
    "A.外卖骑手张某的手机损坏，保管人王某应当赔偿损失",
    "B.外卖骑手张某背部受伤，被救者李某应当适当补偿",
    "C.外卖骑手张某的手机损坏，保管人王某应当适当补偿",
    "D.被救者李某胳膊受伤，外卖骑手张某应当赔偿损失"
  ],
  "answer": "D"
}
```


Ability 7


```
{
  "question": "Delivery rider Zhang saw Li jumping into the river. He handed his phone to a passerby, Wang, and jumped from a height of ten meters into the lake to save Li. During the jump, Zhang accidentally injured his back. Li, unwilling to be rescued, struggled fiercely, causing Zhang to pull harder and worsen his back injury, while Li ended up with a fractured arm. Meanwhile, Wang accidentally dropped Zhang's phone, breaking the screen. Regarding this case, which of the following statements is correct?",
  "options": [
    "A. Wang, the custodian, should compensate for the damage to Zhang's phone.",
    "B. Li, the rescued person, should reasonably compensate Zhang for his back injury.",
    "C. Wang, the custodian, should provide reasonable compensation for the damage to Zhang's phone.",
    "D. Zhang should compensate Li for the arm injury she sustained during the rescue."
  ],
  "answer": "D"
}
```


Figure 15: The MCQ sample of ability 7. The left is the original text, and the right is the English translation.

8. Choosing and correctly citing the relevant laws. Ability to select the most appropriate legal provisions for a given scenario and cite them accurately in reasoning. An MCQ sample of this ability in the LexGenius is shown in Figure 16.

Ability 8


```
{
  "question": "某区市场监督管理局在工作检查时发现某面粉厂房内有外地面粉包装袋，用本地面粉伪造外地面粉进行销售，便扣押了包装袋，扣留了面粉1吨，罚款10万元，该厂不服，提起复议，复议机关区政府复议维持了处罚决定。下列说法正确的是？",
  "options": [
    "A. 监管局对面粉复议鉴定的费用应当由当事人承担",
    "B. 罚款10万元属于法律明确授权范围内的裁量行为",
    "C. 本案被告原则上为区市监局，区政府仅在特定情形下列为共同被告",
    "D. 本案应综合考虑复议机关维持决定的影响来确定级别管辖"
  ],
  "answer": "A"
}
```


Ability 8


```
{
  "question": "During a routine inspection, the Market Supervision Administration of a certain district discovered that a local flour factory was using packaging from other regions to sell local flour as if it were produced elsewhere. The authority confiscated the packaging bags, seized 1 ton of flour, and imposed a fine of 100,000 yuan. ...Upheld the penalty decision. Which of the following statements is correct?",
  "options": [
    "A. The cost of verification and appraisal during reconsideration should be borne by the party concerned.",
    "B. The fine of 100,000 yuan falls within the scope of discretionary power clearly authorized by law.",
    "C. In this case, the defendant is generally the district Market Supervision Administration; the district government is listed as a co-defendant only under specific circumstances.",
    "D. The level of jurisdiction should be determined by comprehensively considering the impact of the reconsideration authority's decision to uphold the original penalty."
  ],
  "answer": "A" }
```


Figure 16: The MCQ sample of ability 8. The left is the original text, and the right is the English translation.

9. Integrate laws across different fields. Ability to synthesize norms from multiple legal domains and resolve inter-norm conflicts through comprehensive analysis. An MCQ sample of this ability in the LexGenius is shown in Figure 17.

Ability 9


```
{
  "question": "某中国公司与法国公司签订了进口某技术产品的贸易合同，后因该技术产品的专利权和海上运输合同产生争议，并诉诸中国某法院，关于相关涉外民事关系，发现《涉外民事关系法律适用法》与其他法律的规定不一致，依我国相关司法解释，下列哪些选项是正确的?( )",
  "options": [
    "A. 均应适用《涉外民事关系法律适用法》的规定",
    "B. 均应适用其他法律的规定",
    "C. 其他法律为《中华人民共和国海商法》的，可由当事人协商决定是否适用《中华人民共和国海商法》的规定",
    "D. 其他法律为知识产权领域的法律的，首先应适用有关知识产权的规定"
  ],
  "answer": "D"
}
```


Ability 9


```
{
  "question": "A Chinese company signed a trade contract with a French company to import a certain technical product. Later, disputes arose over the patent rights of the product and the maritime transport contract, leading to litigation in a Chinese court. Regarding the relevant foreign-related civil relationships, it was found that the provisions of the Law on the Application of Law for Foreign-Related Civil Relationships conflict with other laws. According to relevant judicial interpretations in China, which of the following options is correct?",
  "options": [
    "A. The provisions of the Law on the Application of Law for Foreign-Related Civil Relationships shall apply in all cases.",
    "B. The provisions of other laws shall apply in all cases.",
    "C. If the other law is the Maritime Law of the People's Republic of China, the parties may negotiate to decide whether to apply its provisions.",
    "D. If the other law pertains to intellectual property, the relevant provisions of intellectual property law shall take precedence."
  ],
  "answer": "D" }
```


Figure 17: The MCQ sample of ability 9. The left is the original text, and the right is the English translation.

10. Judging the boundary between law and morality and resolving ethical conflicts. Ability to identify and evaluate tensions between legal obligations and moral principles, and propose ethically aware legal judgments. An MCQ sample of this ability in the LexGenius is shown in Figure 18.

Ability 10

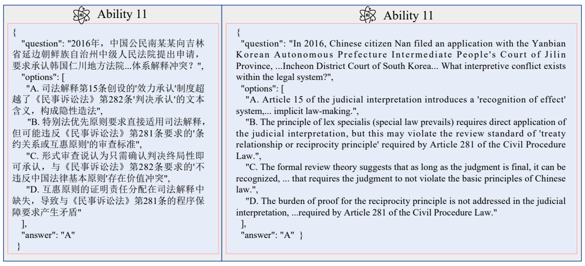

```
{
  "question": "柘城县人口和计划生育委员会对李伟光、李小妞夫妇作出行政征收决定，要求其缴纳社会抚养费。该夫妇因家庭经济困难无力支付，且认为征收标准不合理。在法院准予强制执行后，执行法官发现......在下列选项中应当如何抉择？",
  "options": [
    "A. 限期10日让当事人筹款，逾期仍不履行则启动拍卖程序，期间暂不采取强制措施",
    "B. 立即中止执行程序，协调民政部门启动社会救助，待当事人经济状况改善后恢复执行",
    "C. 严格按生效裁定执行，因法律未明文规定教育权优于征收权，且中止执行可能引发效仿效应",
    "D. 建议计生部门减免全部费用，虽然违反《社会抚养费征收管理办法》但符合人道主义原则"          ],
  "answer": "B"
}
```


Ability 10


```
{
  "question": "The Population and Family Planning Commission of Zhecheng County issued an administrative decision requiring Li Weiguang and his wife, Li Xiaoniu, to pay a social maintenance fee. Due to financial hardship, the couple was unable to pay and believed the fee standard was unreasonable. After the court approved compulsory enforcement, the enforcement judge discovered... What should be the appropriate...?",
  "options": [
    "A. Give the parties a 10-day deadline to raise funds; if they fail to comply, initiate an auction procedure, and refrain ...this period.",
    "B. Immediately suspend the enforcement procedure, coordinate with the civil affairs department to initiate social assistance, and resume enforcement once the parties' financial situation improves.",
    "C. Strictly enforce the effective ruling, since the law does not explicitly ...and suspending enforcement might set a bad precedent.",
    "D. Recommend that the family planning department waive the entire fee, even though this violates ...Social Maintenance Fees, it aligns with humanitarian principles.
"
  ],
  "answer": "B" }
```


Figure 18: The MCQ sample of ability 10. The left is the original text, and the right is the English translation.

11. Critically interpreting legal texts and understanding the lawmakers' intent. Ability to interpret laws beyond their literal wording by uncovering legislative purpose, background, and systemic coherence. An MCQ sample of this ability in the LexGenius is shown in Figure 19.



Ability 11

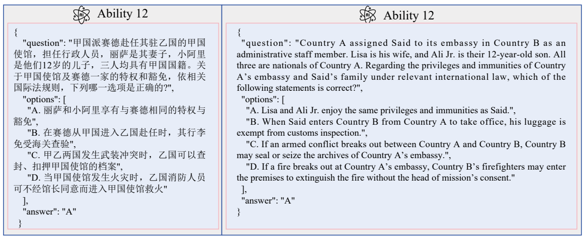
{
"question": "2016年，中国公民南某某向吉林省延边朝鲜族自治州中级人民法院提出申请，要求承认韩国仁川地方法院...体系解释冲突？",
"options": [
"A. 司法解释第15条创设的'效力承认'制度超越了《民事诉讼法》第282条'判决承认'的文本含义，构成隐性造法",
"B. 特别法优先原则要求直接适用司法解释，但可能违反《民事诉讼法》第281条要求的'条约关系或互惠原则'的审查标准",
"C. 形式审查说认为只需确认判决终局性即可承认，与《民事诉讼法》第282条要求的'不违反中国法律基本原则'存在价值冲突",
"D. 互惠原则的证明责任分配在司法解释中缺失，导致与《民事诉讼法》第281条的程序保障要求产生矛盾"
],
"answer": "A"
}

Ability 11

{
"question": "In 2016, Chinese citizen Nan filed an application with the Yanbian Korean Autonomous Prefecture Intermediate People's Court of Jilin Province, ...Incheon District Court of South Korea... What interpretive conflict exists within the legal system?",
"options": [
"A. Article 15 of the judicial interpretation introduces a 'recognition of effect' system,... implicit law-making.",
"B. The principle of lex specialis (special law prevails) requires direct application of the judicial interpretation, but this may violate the review standard of 'treaty relationship or reciprocity principle' required by Article 281 of the Civil Procedure Law.",
"C. The formal review theory suggests that as long as the judgment is final, it can be recognized, ... that requires the judgment to not violate the basic principles of Chinese law.",
"D. The burden of proof for the reciprocity principle is not addressed in the judicial interpretation, ...required by Article 281 of the Civil Procedure Law."
],
"answer": "A" }

Figure 19: The MCQ sample of ability 11. The left is the original text, and the right is the English translation.

12. Interpreting legal terms across fields and adapting to different situations. Ability to understand legal terminology in varied legal contexts and appropriately adapt interpretations to specific domains. An MCQ sample of this ability in the LexGenius is shown in Figure 20.



Ability 12

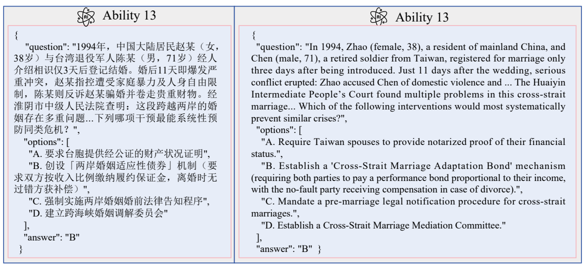
{
"question": "甲国派赛德赴任其驻乙国的甲国使馆，担任行政人员，丽萨是其妻子，小阿里是他们12岁的儿子，三人均具有甲国国籍。关于甲国使馆及赛德一家的特权和豁免，依相关国际法规则，下列哪一选项是正确的?",
"options": [
"A. 丽萨和小阿里享有与赛德相同的特权与豁免",
"B. 在赛德从甲国进入乙国赴任时，其行李免受海关查验",
"C. 甲乙两国发生武装冲突时，乙国可以查封、扣押甲国使馆的档案",
"D. 当甲国使馆发生火灾时，乙国消防人员可不经馆长同意而进入甲国使馆救火"
],
"answer": "A"
}

Ability 12

{
"question": "Country A assigned Said to its embassy in Country B as an administrative staff member. Lisa is his wife, and Ali Jr. is their 12-year-old son. All three are nationals of Country A. Regarding the privileges and immunities of Country A's embassy and Said's family under relevant international law, which of the following statements is correct?",
"options": [
"A. Lisa and Ali Jr. enjoy the same privileges and immunities as Said.",
"B. When Said enters Country B from Country A to take office, his luggage is exempt from customs inspection.",
"C. If an armed conflict breaks out between Country A and Country B, Country B may seal or seize the archives of Country A's embassy.",
"D. If a fire breaks out at Country A's embassy, Country B's firefighters may enter the premises to extinguish the fire without the head of mission's consent."
],
"answer": "A"
}

Figure 20: The MCQ sample of ability 12. The left is the original text, and the right is the English translation.

13. Understanding the exact meaning of legal terms. Ability to grasp the technical definitions, scope, and usage boundaries of domain-specific legal terms. An MCQ sample of this ability in the LexGenius is shown in Figure 21.



Ability 13

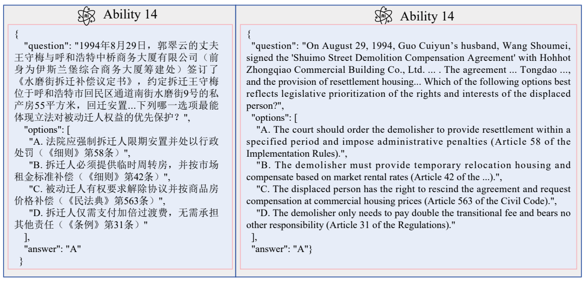
{
"question": "1994年，中国大陆居民赵某（女，38岁）与台湾退役军人陈某（男，71岁）经人介绍相识仅3天后登记结婚。婚后11天即爆发严重冲突，赵某指控遭受家庭暴力及人身自由限制，陈某则反诉赵某骗婚并卷走贵重财物。经淮阴市中级人民法院查明：这段跨越两岸的婚姻存在多重问题...下列哪项干预最能系统性预防同类危机？",
"options": [
"A. 要求台胞提供经公证的财产状况证明",
"B. 创设「两岸婚姻适应性债券」机制（要求双方按收入比例缴纳履约保证金，离婚时无过错方获补偿）",
"C. 强制实施两岸婚姻婚前法律告知程序",
"D. 建立跨海峡婚姻调解委员会"
],
"answer": "B"
}

Ability 13

{
"question": "In 1994, Zhao (female, 38), a resident of mainland China, and Chen (male, 71), a retired soldier from Taiwan, registered for marriage only three days after being introduced. Just 11 days after the wedding, serious conflict erupted: Zhao accused Chen of domestic violence and ... The Huaiyin Intermediate People's Court found multiple problems in this cross-strait marriage... Which of the following interventions would most systematically prevent similar crises?",
"options": [
"A. Require Taiwan spouses to provide notarized proof of their financial status.",
"B. Establish a 'Cross-Strait Marriage Adaptation Bond' mechanism (requiring both parties to pay a performance bond proportional to their income, with the no-fault party receiving compensation in case of divorce).",
"C. Mandate a pre-marriage legal notification procedure for cross-strait marriages.",
"D. Establish a Cross-Strait Marriage Mediation Committee."
],
"answer": "B" }

Figure 21: The MCQ sample of ability 13. The left is the original text, and the right is the English translation.

14. Analyzing the social impact and stability of legal enforcement. Ability to assess the potential impact of legal implementation on public order, institutional trust, and long-term societal effects. An MCQ sample of this ability in the LexGenius is shown in Figure 22.



Ability 14

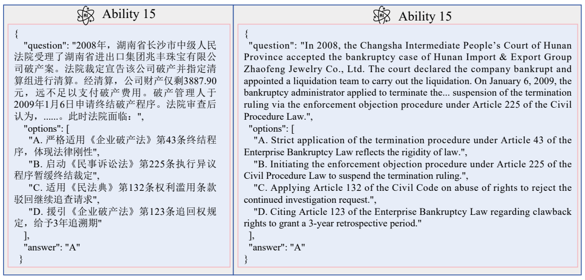
{
"question": "1994年8月29日，郭翠云的丈夫王守梅与呼和浩特中桥商务大厦有限公司（前身为伊斯兰堡综合商务大厦筹建处）签订了《水磨街拆迁补偿议定书》，约定拆迁王守梅位于呼和浩特市回民区通道南街水磨街9号的私产房55平方米，回迁安置...下列哪一选项最能体现立法对被动迁人权益的优先保护？",
"options": [
"A. 法院应强制拆迁人限期安置并处以行政处罚（《细则》第58条）",
"B. 拆迁人必须提供临时周转房，并按市场租金标准补偿（《细则》第42条）",
"C. 被动迁人有权要求解除协议并按商品房价格补偿（《民法典》第563条）",
"D. 拆迁人仅需支付加倍过渡费，无需承担其他责任（《条例》第31条）"
],
"answer": "A"
}

Ability 14

{
"question": "On August 29, 1994, Guo Cuiyun's husband, Wang Shoumei, signed the 'Shuimo Street Demolition Compensation Agreement' with Hohhot Zhongqiao Commercial Building Co., Ltd. ... . The agreement ... Tongdao ..., and the provision of resettlement housing... Which of the following options best reflects legislative prioritization of the rights and interests of the displaced person?",
"options": [
"A. The court should order the demolisher to provide resettlement within a specified period and impose administrative penalties (Article 58 of the Implementation Rules).",
"B. The demolisher must provide temporary relocation housing and compensate based on market rental rates (Article 42 of the ...).",
"C. The displaced person has the right to rescind the agreement and request compensation at commercial housing prices (Article 563 of the Civil Code).",
"D. The demolisher only needs to pay double the transitional fee and bears no other responsibility (Article 31 of the Regulations)."
],
"answer": "A"}

Figure 22: The MCQ sample of ability 14. The left is the original text, and the right is the English translation.

15. Social change, culture, and legal coordination. Ability to understand how law responds to social transformation and interacts with culture, economy, and values. An MCQ sample of this ability in the LexGenius is shown in Figure 23.



Ability 15

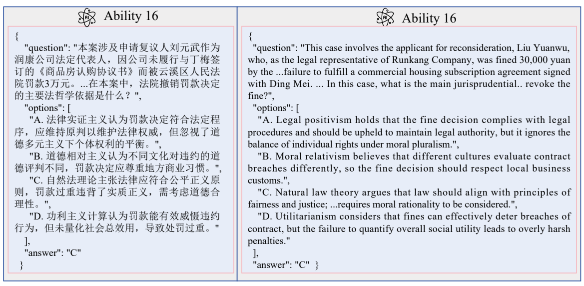
{
"question": "2008年，湖南省长沙市中级人民法院受理了湖南省进出口集团兆丰珠宝有限公司破产案。法院裁定宣告该公司破产并指定清算组进行清算。经清算，公司财产仅剩3887.90元，远不足以支付破产费用。破产管理人于2009年1月6日申请终结破产程序。法院审查后认为，......。此时法院面临：",
"options": [
"A. 严格适用《企业破产法》第43条终结程序，体现法律刚性",
"B. 启动《民事诉讼法》第225条执行异议程序暂缓终结裁定",
"C. 适用《民法典》第132条权利滥用条款驳回继续追查请求",
"D. 援引《企业破产法》第123条追回权规定，给予3年追溯期"
],
"answer": "A"
}

Ability 15

{
"question": "In 2008, the Changsha Intermediate People's Court of Hunan Province accepted the bankruptcy case of Hunan Import & Export Group Zhaofeng Jewelry Co., Ltd. The court declared the company bankrupt and appointed a liquidation team to carry out the liquidation. On January 6, 2009, the bankruptcy administrator applied to terminate the... suspension of the termination ruling via the enforcement objection procedure under Article 225 of the Civil Procedure Law.",
"options": [
"A. Strict application of the termination procedure under Article 43 of the Enterprise Bankruptcy Law reflects the rigidity of law.",
"B. Initiating the enforcement objection procedure under Article 225 of the Civil Procedure Law to suspend the termination ruling.",
"C. Applying Article 132 of the Civil Code on abuse of rights to reject the continued investigation request.",
"D. Citing Article 123 of the Enterprise Bankruptcy Law regarding clawback rights to grant a 3-year retrospective period."
],
"answer": "A"
}

Figure 23: The MCQ sample of ability 15. The left is the original text, and the right is the English translation.

16. Understanding and managing conflicts between law and morality. Ability to propose socially responsible legal judgments in situations where legal and moral norms collide. An MCQ sample of this ability in the LexGenius is shown in Figure 24.



Ability 16

{
"question": "本案涉及申请复议人刘元武作为润康公司法定代表人，因公司未履行与丁梅签订的《商品房认购协议书》而被云溪区人民法院罚款3万元。...在本案中，法院撤销罚款决定的主要法哲学依据是什么？",
"options": [
"A. 法律实证主义认为罚款决定符合法定程序，应维持原判以维护法律权威，但忽视了道德多元主义下个体权利的平衡。",
"B. 道德相对主义认为不同文化对违约的道德评判不同，罚款决定应尊重地方商业习惯。",
"C. 自然法理论主张法律应符合公平正义原则，罚款过重违背了实质正义，需考虑道德合理性。",
"D. 功利主义计算认为罚款能有效威慑违约行为，但未量化社会总效用，导致处罚过重。"
],
"answer": "C"
}

Ability 16

{
"question": "This case involves the applicant for reconsideration, Liu Yuanwu, who, as the legal representative of Runkang Company, was fined 30,000 yuan by the ...failure to fulfill a commercial housing subscription agreement signed with Ding Mei. ... In this case, what is the main jurisprudential.. revoke the fine?",
"options": [
"A. Legal positivism holds that the fine decision complies with legal procedures and should be upheld to maintain legal authority, but it ignores the balance of individual rights under moral pluralism.",
"B. Moral relativism believes that different cultures evaluate contract breaches differently, so the fine decision should respect local business customs.",
"C. Natural law theory argues that law should align with principles of fairness and justice; ...requires moral rationality to be considered.",
"D. Utilitarianism considers that fines can effectively deter breaches of contract, but the failure to quantify overall social utility leads to overly harsh penalties."
],
"answer": "C" }

Figure 24: The MCQ sample of ability 16. The left is the original text, and the right is the English translation.

17. Reasonable legal reasoning and judgment prediction under uncertainty. Ability to make legally sound decisions when faced with ambiguous facts or normative gaps, using analogical reasoning and proportionality. An MCQ sample of this ability in the LexGenius is shown in Figure 25.

Ability 17

{
"question": "许某与妻子林某协议离婚，约定8岁的儿子小虎由许某抚养，林某可随时行使对儿子的探望权，许某有协助的义务。离婚后两年间林某从未探望过儿子，小虎诉至法院，要求判令林某每月探视自己不少于4天。对此，下列说法正确的是：",
"options": [
"A. 根据法律规定，探望权的行使可通过司法判决强制实施，以体现子女最大利益原则",
"B. 权利的行使与义务的履行虽有界限，但在涉及未成年子女权益时可适当突破该界限",
"C. 探望权本质上是一种附带作为义务的权利，权利人需依规定主动履行相应行为",
"D. 许某的协助义务同时包括积极义务和消极义务"
],
"answer": "D"
}

Ability 17

{
"question": "Xu and his wife Lin agreed to divorce, with the arrangement that their 8-year-old son Xiaohu would be raised by Xu, and Lin could exercise her visitation rights at any time, with Xu obliged to assist. In the two years following the divorce, ... Regarding this situation, which of the following statements is correct?",
"options": [
"A. According to the law, the exercise of visitation rights can be enforced through judicial ruling to uphold the best interests of the child.",
"B. Although there are boundaries between the exercise of rights and the fulfillment of obligations, such boundaries may be appropriately adjusted when the rights of minors are involved.",
"C. Visitation rights are essentially rights with attached positive obligations, and the right-holder must proactively fulfill the required acts.",
"D. Xu's duty to assist includes both active and passive obligations."
],
"answer": "D" }

Figure 25: The MCQ sample of ability 17. The left is the original text, and the right is the English translation.

18. Case-based reasoning and judgment. Ability to construct judgments through analogical reasoning with relevant precedents and case-specific facts. An MCQ sample of this ability in the LexGenius is shown in Figure 26.

Ability 18

{
"question": "1997年7月起，刘承明未经审批在青岛市城阳区河套镇山角村渔港码头违法经营加油站。1999年5月9日，刘承明为'鲁城渔0586'号渔船加油时，因所售柴油不合格（闭口闪点27℃低于国家标准65℃且含汽油成分）、加油器具未作防静电处理...此时法官应如何认定柴油的法律属性？",
"options": [
"A. 中止审理并层报最高法院，待出台明确司法解释后再行判决",
"B. 采用体系解释方法，结合农业部规定中'油气积聚危险'的立法目的，将本案柴油扩张解释为刑法第136条的'易燃物品'",
"C. 根据'存疑有利于被告人'原则，以法律未明文规定为由否定柴油的易燃性，宣告被告人无罪",
"D. 严格遵循公安部文件的技术标准，认定闪点27℃的柴油不符合易燃物品定义，直接否定犯罪构成要件"],
"answer": "B" }

Ability 18

{
"question": "Starting in July 1997, Liu Chengming illegally operated a gas station without approval at the fishing port dock of Shanjiao Village, Hetao Town,... On May 9, 1999, while refueling the fishing boat 'Lu Cheng Yu 0586', the diesel fuel sold by Liu was substandard ..., and the refueling equipment lacked anti-static treatment...How should the judge determine the legal classification of the diesel?",
"options": [
"A. Suspend the trial and report to the Supreme People's Court, waiting for a clear judicial interpretation before rendering a verdict.",
"B. Apply systematic interpretation, and in light of the legislative intent behind ..., expansively interpret the diesel in this case as 'flammable material' under Article 136 of the Criminal Law.",
"C. Apply the principle of 'in dubio pro reo' , and deny the flammability of diesel due to the lack of explicit legal definition, declaring the defendant not guilty.",
"D. Strictly follow the technical ..., and determine that diesel with a flash point of 27°C ..., thus negating the elements of the crime."
],
"answer": "B" }

Figure 26: The MCQ sample of ability 18. The left is the original text, and the right is the English translation.

19. Analysis of the application of judicial procedures in different jurisdictions. Ability to identify jurisdictional differences in judicial procedures and adjust legal reasoning accordingly. An MCQ sample of this ability in the LexGenius is shown in Figure 27.

Ability 19

{
"question": "1999年3月15日晚8时30分许，被告人赵剑峰驾驶丹东市浪头造纸机械厂的无牌照面包车，在振兴区人民路由东向西行驶。当行至市第二医院附近路口时，将横过马路的行人裴凤金撞倒。事故发生后，赵剑峰与同车人邹某、王某将...下列关于二审程序的说法哪项是正确的？",
"options": [
"A. 二审法院应当对案件进行全面审理，不受上诉范围限制，但不得加重被告人刑罚",
"B. 赵剑峰的上诉期限从接到判决书之日起算，但因交通不便可向法院申请延长至15日",
"C. 赵剑峰必须通过原审法院提出上诉，且上诉状副本应直接送达同级人民检察院，否则视为无效上诉",
"D. 若赵剑峰仅就民事赔偿部分提出上诉，二审法院仍有权对刑事部分进行审查并改判"
],
"answer": "A"
}

Ability 19

{
"question": "On the evening of March 15, 1999, around 8:30 PM, defendant Zhao Jianfeng was driving an unlicensed van belonging to the Langtou Paper Machinery Factory in Dandong. While heading west on Renmin Road in Zhenxing District near the Second City Hospital, he hit a pedestrian, Pei Fengjin. After the accident, Zhao and two passengers, Zou and Wang, ... Which of the following statements about the second-instance procedure is correct?",
"options": [
"A. The second-instance court shall conduct a full ... of the appeal, but shall not impose a heavier sentence on the defendant.",
"B. Zhao Jianfeng's appeal period starts from the day he receives the judgment, and may be extended to 15 days due to transportation difficulties upon request.",
"C. Zhao must submit the appeal through the original court, and the appeal copy ... otherwise the appeal is invalid.",
"D. If Zhao only appeals the civil compensation part, the second-instance court may still review and revise the criminal part."
],
"answer": "A" }

Figure 27: The MCQ sample of ability 19. The left is the original text, and the right is the English translation.

20. Understanding of judicial procedures and the ability to grasp details. Ability to accurately apply procedural rules throughout litigation or non-litigation processes, ensuring procedural compliance. An MCQ sample of this ability in the LexGenius is shown in Figure 28.

Ability 20

{
"question": "互联网法院在适用普通程序审理案件过程中，下列表述错误的有：",
"options": [
"A. 互联网法院采取在线方式审理案件，诉讼过程中发现某诉讼环节无法通过线上完成的，应当裁定中止审理，待线下完成该环节后，再恢复该案的线上处理程序",
"B. 互联网法院在线接收并处理当事人提交的诉讼材料，处理结果形成电子文件并通过网络送达当事人",
"C. 互联网法院采取在线视频方式开庭、组织证据交换，并可依具体情况决定是否需要线下补充质证",
"D. 对于简单民事案件，互联网法院可以依当事人申请将当事人陈述、法庭调查、法庭辩论等庭审环节合并进行"
],
"answer": "A"
}

Ability 20

{
"question": "In the course of using ordinary procedures to hear cases, which of the following statements about Internet Courts is incorrect?",
"options": [
"A. If any litigation step cannot be completed online during trial, the Internet Court must rule to suspend the trial and resume it only after the step is completed offline.",
"B. The Internet Court accepts and processes litigation materials submitted online, and delivers the results as electronic documents via the internet.",
"C. The Internet Court conducts hearings and evidence exchange via online video, and may decide whether offline cross-examination is needed based on specific circumstances.",
"D. For simple civil cases, the Internet Court may, upon a party's request, merge the steps of statement, investigation, and debate into one hearing session."
],
"answer": "A"
}

Figure 28: The MCQ sample of ability 20. The left is the original text, and the right is the English translation.

C Annotation Details

We recruited nine master's candidates in law for double-blind annotation. As detailed in Section 4.2 and Figure 3, evaluators assessed questions on a 5-point scale based on five specific criteria: statutory accuracy, logical soundness, answer uniqueness, competence alignment, and normative wording. Arbitration was triggered for score discrepancies exceeding two points. Participants were informed of data usage and management policies. To ensure fair labor practices, annotators were compensated at a rate of US$15 per hour, which exceeds the local average wage for student research assistants.

D Experimental setup details

D.1 Large Language Models

We evaluated 12 LLMs on LexGenius. For GPT-4o mini, GPT-4 nano, DeepSeek-V3, and DeepSeek-R1, we accessed them via their official APIs. For other LLMs, we conducted experimental tests using the official weights. These LLMs are as follows:

Qwen2.5-1.5B-Instruct (Hui et al., 2024): A lightweight instruction-tuned model released by Alibaba with 1.5B parameters, designed for edge deployment and local inference with bilingual support and basic task execution.

Qwen2.5-7B-Instruct (Hui et al., 2024): A mid-scale model in the Qwen2.5 series, optimized for stronger reasoning and instruction following, suitable for more complex language tasks in medium-sized deployments.

Qwen3-4B (Yang et al., 2025): A 4B-parameter model from the third-generation Qwen series, showing strong performance in multilingual, coding, and logical tasks.

Qwen3-8B (Yang et al., 2025): An enhanced version of Qwen3 with extended context length and multilingual capabilities, significantly improving performance in reasoning and generation tasks.

LLaMA-3.2-1B-Instruct (Grattafiori et al., 2024): A compact instruction-tuned model from Meta's LLaMA 3 series, designed for resource-constrained environments while maintaining core instruction-following capabilities.

LLaMA-3.2-8B-Instruct (Grattafiori et al., 2024): A standard model in the LLaMA 3 lineup, offering high-quality multilingual understanding, code generation, and reasoning, achieving state-of-the-art results across many tasks.

GLM-4-9B-Chat (GLM et al., 2024): A 9B bilingual chat model developed by Zhipu AI, equipped for multi-turn dialogue, tool use, and contextual memory, with strong performance particularly in Chinese semantic understanding.

DeepSeek-LLM-7B-Chat (Bi et al., 2024): A 7B bilingual chat model by DeepSeek, integrating capabilities in code generation, mathematics, and language understanding, suitable for dialogue and multitask settings.

DeepSeek-R1 (Guo et al., 2025): A large-scale open-source language model developed by the Chinese company DeepSeek, featuring strong capabilities in mathematics, programming, and reasoning with efficient training and leading performance.

DeepSeek-V3 (Liu et al., 2024): A large language model based on a mixture-of-experts architecture, excelling in mathematics, programming, and logical reasoning, and is well-suited for a variety of intelligent application scenarios.

GPT-4o mini (Hurst et al., 2024): A parameter-efficient version of OpenAI's GPT-4o, supporting multimodal inputs (text, image, audio) with consistent alignment behavior as its full-size counterpart.

GPT-4.1 nano (Hurst et al., 2024): An ultra-compact model in the GPT-4.1 series, designed for on-device and embedded inference with support for moderate-length contexts and basic reasoning under resource constraints.

D.2 Two Prompt Methods

To evaluate LexGenius, we employed Naive and CoT strategies (see Figure 29). The Naive prompt prioritizes efficiency via direct output generated in a single pass, though its lack of reasoning often degrades performance on complex tasks. Conversely, CoT simulates human problem-solving by decomposing tasks into intermediate steps, activating causal chains to significantly reduce errors in multi-step dependencies and conditional logic.

Naive Prompt (Original text): ""假如你是一个专业的法律人员，请根据你的法律知识，在A，B，C，D四个选项中选择1个正确的选项，只回答A或B或C或D四个字母中的一个，只回答正确选项的序号（即A或B或C或D）不要回答其他的内容",

Naive Prompt (English translation): "If you are a professional legal professional, please choose the correct option from the four options, A, B, C, and D, based on your legal knowledge. Only answer one of the four letters, A, B, C, or D. Only answer the sequence number of the correct option (i.e., A, B, C, or D). Do not answer anything else.",

CoT Prompt (Original text): "假如你是一个专业的法律人员，请根据你的法律知识，进行一步一步的深入思考，在A，B，C，D四个选项中选择1个正确的选项，只回答A或B或C或D四个字母中的一个，只回答正确选项的序号（即A或B或C或D）不要回答其他的内容",

CoT Prompt (English translation): "If you are a professional legal professional, please think deeply and step by step based on your legal knowledge. Choose one correct option from the four options A, B, C, and D. Only answer one of the four letters, A, B, C, or D. Only answer with the number of the correct option (i.e., A, B, C, or D). Do not answer anything else.",

Figure 29: The two utilized prompt methods for LLMs. In this figure, we provide the Chinese and English texts.

E Results of Twenty Abilities

Based on the structured capability framework provided by LexGenius, we evaluated the performance of various SOTA LLMs across 20 legal intelligence abilities (see Table 7). LexGenius is categorized under 7 core dimensions. These dimensions are further divided into 11 tasks and 20 abilities, covering a comprehensive range of legal intelligence abilities, including understanding, reasoning, application, ethical judgment, language processing, socio-legal interaction, and judicial practice.

The results of 20 legal intelligence abilities for the 12 LLMs are shown in Table 7. The LLMs' legal intelligence abilities decline significantly in tasks requiring deeper abstraction. These tasks involve complex value judgments, cross-domain norm integration, and procedural reasoning—areas where LLMs struggle to match human-like legal cognition. This highlights the need for further model optimization in sociological, ethical, and institutional aspects of legal general intelligence.

F With Different Enhanced Methods

To evaluate the impact of different optimization and enhancement methods on the legal intelligence capabilities of LLMs, we selected four LLMs (including Qwen2.5-1.5B-Instruct, Qwen2.5-7B-Instruct, Qwen3-4B, and Qwen3-8B) and experimented with Supervised Fine-Tuning (SFT), Chain-of-Thought (CoT), Retrieval-Augmented Generation (RAG), and Reinforcement Learning (RL) algorithms. We randomly sampled 64 test instances from each of the 20 ability test sets in LexGenius, resulting in 1,280 total samples for evaluation. The remaining 7,105 data samples were used as the training set for SFT and RL, as well as for construct-

Ability	Human	LLM 1	LLM 2	LLM 3	LLM 4	LLM 5	LLM 6	LLM 7	LLM 8	LLM 9	LLM 10	LLM 11	LLM 12
Naive Prompt													
Ability 1	89.63	45.00	58.20	50.80	55.20	28.80	37.80	53.00	39.00	64.40	64.80	47.00	41.40
Ability 2	87.41	37.78	60.89	46.22	54.22	30.67	33.33	46.22	34.22	67.11	65.78	47.11	40.44
Ability 3	90.37	49.38	65.15	62.66	65.15	28.63	39.83	59.34	38.59	70.54	70.54	55.19	57.68
Ability 4	92.59	39.40	56.60	48.00	50.60	25.80	32.20	47.60	32.20	59.20	58.00	46.00	39.00
Ability 5	88.89	43.45	55.36	52.38	50.60	23.81	27.98	41.67	35.71	60.71	60.12	39.29	47.62
Ability 6	85.93	40.19	60.29	50.72	55.02	22.49	33.97	54.07	34.93	63.64	64.59	44.50	43.06
Ability 7	85.93	36.20	47.80	41.40	42.80	25.60	31.60	39.40	35.80	46.80	46.80	40.40	35.40
Ability 8	92.59	39.60	52.20	46.00	48.40	23.20	34.80	42.60	27.40	53.00	55.20	40.40	35.80
Ability 9	88.89	31.03	35.78	29.74	25.00	30.60	33.19	24.57	25.43	37.93	37.07	30.17	21.55
Ability 10	86.67	59.00	64.80	58.00	64.20	46.80	56.20	66.00	48.40	67.60	67.60	65.20	59.20
Ability 11	83.70	52.40	63.20	63.80	61.60	42.00	50.80	56.80	41.80	74.00	73.00	59.20	54.40
Ability 12	89.63	43.55	63.55	50.65	60.97	27.74	29.35	58.71	32.26	74.19	73.55	45.81	45.81
Ability 13	85.93	60.40	68.60	64.60	65.20	51.00	63.40	66.20	46.80	68.00	67.40	69.20	60.40
Ability 14	87.41	57.20	71.00	69.00	67.60	45.80	55.00	64.40	51.60	76.80	76.40	67.60	63.00
Ability 15	91.85	56.40	58.00	55.60	54.80	46.00	51.60	57.60	40.20	65.40	66.20	61.20	54.40
Ability 16	82.96	64.40	69.20	62.00	61.20	59.00	59.60	63.60	49.20	75.00	72.60	67.60	59.00
Ability 17	89.63	38.40	45.80	40.80	45.00	23.80	34.20	39.40	33.40	48.80	49.40	40.20	37.00
Ability 18	90.37	50.80	64.60	58.20	66.00	44.00	53.60	60.00	40.40	72.40	71.80	60.40	59.00
Ability 19	79.26	58.20	74.00	65.80	69.60	37.80	54.60	63.80	45.60	80.00	81.20	65.80	61.80
Ability 20	87.41	34.60	45.80	41.40	40.80	26.60	28.80	39.00	29.40	48.40	47.20	42.60	35.00
CoT Prompt													
Ability 1	89.63	45.00	57.60	50.60	54.40	30.00	36.40	53.20	36.80	64.40	64.20	48.00	56.80
Ability 2	87.41	38.67	60.00	44.44	55.11	28.44	32.00	45.78	30.67	66.67	66.67	46.22	50.22
Ability 3	90.37	48.13	63.90	62.24	66.80	26.56	38.59	60.17	33.20	72.20	71.78	55.19	56.85
Ability 4	92.59	42.00	56.20	48.00	49.20	23.60	32.20	47.00	32.00	60.40	58.80	45.80	40.00
Ability 5	88.89	42.26	54.17	52.98	50.60	28.57	30.95	41.07	27.38	59.52	61.31	40.48	47.62
Ability 6	85.93	42.58	59.33	50.72	58.37	25.36	35.41	54.55	33.49	64.59	64.59	45.45	44.02
Ability 7	85.93	36.00	48.00	42.00	41.40	24.60	32.00	39.60	34.40	46.60	47.60	39.40	39.40
Ability 8	92.59	39.60	51.80	44.20	47.40	22.20	35.00	41.80	29.80	53.60	53.40	40.60	48.00
Ability 9	88.89	31.90	35.78	28.02	22.84	31.90	33.19	28.02	28.45	37.50	38.79	29.31	32.76
Ability 10	86.67	59.40	65.40	58.60	62.80	46.60	55.20	65.60	50.20	68.00	67.40	65.80	60.80
Ability 11	83.70	52.60	61.80	63.40	60.40	40.20	51.00	57.20	36.40	73.20	71.40	58.60	58.40
Ability 12	89.63	44.84	63.23	50.32	60.65	26.45	29.03	59.68	32.58	74.19	73.23	46.45	60.97
Ability 13	85.93	61.80	68.60	64.40	63.20	53.40	64.20	66.80	49.00	68.20	67.80	68.60	64.40
Ability 14	87.41	57.20	70.40	68.20	66.00	41.40	56.40	63.40	50.40	77.00	77.20	67.20	66.20
Ability 15	91.85	56.20	59.60	54.00	54.60	45.60	51.80	57.40	42.60	65.00	65.20	61.40	54.00
Ability 16	82.96	66.00	68.80	61.20	60.80	56.80	61.20	63.60	49.20	75.00	75.40	68.80	61.40
Ability 17	89.63	38.80	45.20	41.20	44.40	21.60	32.40	38.40	30.60	51.20	48.80	39.60	37.00
Ability 18	90.37	50.80	64.60	58.00	65.00	42.80	54.80	60.80	42.60	73.60	72.80	61.80	60.60
Ability 19	79.26	59.20	74.40	66.00	68.40	39.00	56.60	64.60	43.60	79.60	81.00	67.00	66.00
Ability 20	87.41	36.60	46.20	41.80	39.80	25.60	30.40	40.40	29.40	47.00	47.40	42.00	46.80

Table 7: Comparison of performance across 20 legal intelligence abilities for Naive Prompt and CoT Prompt on various LLMs (all values in %). LLM 1 is Qwen2.5-1.5B-Instruct; LLM 2 is Qwen2.5-7B-Instruct; LLM 3 is Qwen3-4B; LLM 4 is Qwen3-8B; LLM 5 is Llama-3.2-1B-Instruct; LLM 6 is Llama-3.2-8B-Instruct; LLM 7 is GLM-4-9B-Chat; LLM 8 is DeepSeek-LLM-7B-Chat; LLM 9 is DeepSeek-R1; LLM 10 is DeepSeek-V3; LLM 11 is GPT-4o mini; and LLM 12 is GPT-4.1 nano.

ing the retrieval corpus. The appropriate parameters were selected for SFT and RL training. The experimental results of these LLMs, after applying these enhancement methods across various dimensions and tasks of legal intelligence, are shown in Table 8 and Table 9.

G Correlation Analysis

The performance of 12 LLMs on LexGenius is utilized to analyze correlations (see Figure 30). It illustrates that most of the legal intelligence abilities (left), tasks (upper right), and dimensions (lower right) exhibit low correlations. It shows the effectiveness of LexGenius because the low intercorre-

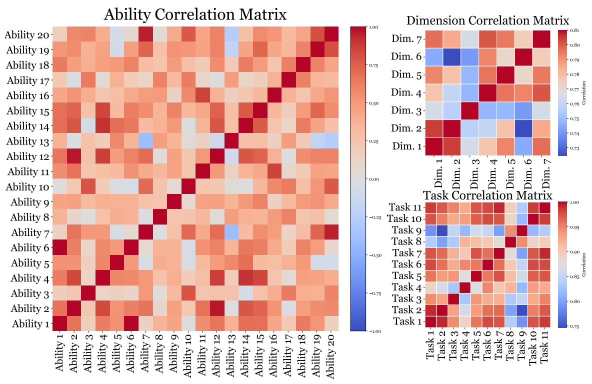


Figure 30: The correlation analysis of legal intelligence ability, task, and dimension in LexGenius for 12 LLMs.

Model	Legal Und.	Legal Rea.	Legal App.	Legal Ethics	Legal Lan.	Law & Soc.	Judicial Pra.	Avg.
				Baseline				
Qwen3-4B	22.39	31.77	32.82	30.47	18.75	33.85	35.16	29.32
Qwen2.5-7B	**62.50**	51.04	**42.19**	61.72	63.28	69.79	**58.60**	**58.45**
Qwen3-8B	25.52	31.25	29.69	27.34	17.97	32.81	30.47	27.86
Qwen2.5-1.5B	43.75	44.79	35.41	56.25	51.57	58.33	49.61	48.53
				CoT				
Qwen3-4B	19.79	29.69	34.38	29.69	18.75	34.37	32.03	28.39
Qwen2.5-7B	60.94	52.09	38.54	60.16	61.72	69.27	57.42	57.16
Qwen3-8B	23.43	29.69	33.33	28.91	19.53	30.73	28.91	27.79
Qwen2.5-1.5B	42.19	46.35	35.42	57.82	52.35	61.98	50.00	49.44
				RAG				
Qwen3-4B	37.67	35.64	34.43	36.86	36.35	42.70	40.02	37.67
Qwen2.5-7B	57.29	45.84	41.14	53.13	54.89	57.77	55.47	52.22
Qwen3-8B	48.96	47.92	28.64	51.56	48.44	50.00	44.93	45.78
Qwen2.5-1.5B	30.93	34.54	32.08	37.50	32.81	38.71	36.06	34.66
				SFT				
Qwen3-4B	50.52	37.50	33.86	67.19	52.35	67.71	50.00	51.30
Qwen2.5-7B	56.25	52.08	31.25	64.85	61.72	66.66	58.20	55.86
Qwen3-8B	61.98	45.83	32.81	**69.53**	63.28	**71.35**	53.13	56.84
Qwen2.5-1.5B	49.48	43.75	31.25	60.16	61.72	67.19	46.87	51.49
				GRPO				
Qwen3-4B	52.08	47.92	29.17	60.94	53.91	63.02	50.00	51.01
Qwen2.5-7B	57.29	**53.65**	35.42	56.25	**67.97**	63.54	57.42	55.93
Qwen3-8B	59.90	50.00	35.94	60.94	63.28	62.50	52.74	55.04
Qwen2.5-1.5B	53.65	46.36	32.81	62.50	60.94	61.46	49.61	52.48

Table 8: Comparison of the four LLMs with different enhanced methods on seven dimensions of LexGenius, which include CoT, RAG, SFT, and GRPO.

Model	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Avg.
						Baseline						
Qwen3-4B	22.39	31.77	32.82	28.12	32.81	18.75	37.50	39.06	25.00	30.46	39.84	30.78
Qwen2.5-7B	**62.50**	51.04	**42.19**	59.38	**64.06**	63.28	**68.75**	67.19	73.44	53.91	**63.28**	**60.82**
Qwen3-8B	25.52	31.25	29.69	28.12	26.56	17.97	35.94	37.50	25.00	27.34	33.59	28.95
Qwen2.5-1.5B	43.75	44.79	35.41	57.81	54.69	51.56	53.12	60.94	60.94	50.00	49.22	51.11
						CoT						
Qwen3-4B	19.79	29.69	34.38	28.12	31.25	18.75	40.62	37.50	25.00	28.91	35.16	29.92
Qwen2.5-7B	60.94	52.09	38.54	56.25	**64.06**	61.72	67.19	67.19	73.44	52.34	62.50	59.66
Qwen3-8B	23.43	29.69	33.33	29.69	28.12	19.53	29.69	39.06	23.44	25.00	32.81	28.53
Qwen2.5-1.5B	42.19	46.35	35.42	59.38	56.25	52.34	53.12	65.62	67.19	51.56	48.44	52.53
						RAG						
Qwen3-4B	37.67	35.64	34.43	38.71	35.00	36.35	54.84	30.65	42.62	40.05	39.98	38.72
Qwen2.5-1.5B	30.93	34.54	32.08	40.62	34.38	32.81	33.33	50.00	32.81	38.28	33.84	35.78
Qwen3-8B	48.96	47.92	28.64	53.12	50.00	48.44	53.12	43.75	53.12	42.19	47.66	46.99
Qwen2.5-7B	57.29	45.84	41.14	56.25	50.00	54.89	55.56	58.06	59.68	50.78	60.16	53.60
						SFT						
Qwen3-4B	50.52	37.50	33.86	73.44	60.94	52.34	60.94	**70.31**	71.88	46.09	53.91	55.61
Qwen2.5-7B	56.25	52.08	31.25	70.31	59.38	61.72	65.62	65.62	68.75	59.38	57.03	58.85
Qwen3-8B	61.98	45.83	32.81	**75.00**	**64.06**	63.28	64.06	68.75	**81.25**	53.91	52.34	60.30
Qwen2.5-1.5B	49.48	43.75	31.25	70.31	50.00	61.72	64.06	67.19	70.31	48.44	45.31	54.71
						GRPO						
Qwen3-4B	52.08	47.92	29.17	62.50	59.38	53.91	67.19	56.25	65.62	49.22	50.78	54.00
Qwen2.5-7B	57.29	**53.65**	35.42	57.81	54.69	**67.97**	62.50	60.94	67.19	**61.72**	53.12	57.48
Qwen3-8B	59.90	50.00	35.94	60.94	60.94	63.28	57.81	59.38	70.31	55.47	50.00	56.72
Qwen2.5-1.5B	53.65	46.36	32.81	68.75	56.25	60.94	62.50	56.25	65.62	52.34	46.87	54.76

Table 9: Comparison of the four LLMs with different enhanced methods on eleven tasks of LexGenius, which include CoT, RAG, SFT, and GRPO.

lation suggests LLMs cannot rely on general legal heuristics or shallow transfer across domains to perform well; instead, success in one category does not guarantee success in others. This reflects the comprehensive coverage and conceptual independence of our benchmark dimensions, further validating their robustness as an evaluation framework.

H Limitations

Although LexGenius structures legal general intelligence evaluation, it is limited by a lack of multimodal capabilities, cross-jurisdictional coverage, and temporal awareness. These gaps constrain its ability to capture real-world complexity. The following subsections detail these core limitations.

Lack of Multimodal Tasks Limits Realistic Evidence Modeling. The current version of LexGenius relies entirely on pure textual materials, excluding multimodal evidence types common in real-world cases, such as scanned contracts, video stills, or audio transcriptions. This unimodal design fails to assess capabilities in visual perception, auditory understanding, and cross-modal reasoning essential for handling actual judicial cases. Consequently, the absence of multimodal inputs limits LLM applicability in tasks such as evidence review, fact reconstruction, and visual-legal interpretation, reducing evaluation fidelity to real-world scenarios.

Linguistic and Jurisdictional Limitations Undermine Cross-Cultural Generalization. LexGenius is currently constructed solely from Chinese corpora and Mainland China's law system, exhibiting distinct linguistic and legal singularity. Consequently, evaluations are confined to this context, failing to capture broader capabilities like interpreting international statutes or comparative analysis. This restriction limits applicability in global legal services and cross-border disputes, hampering transferability and impeding the evolution into a universal legal intelligence system.

Lack of Evaluation on Temporal Sensitivity and Legal Validity Awareness. A core characteristic of law is its temporal nature. Applicable rules for a given issue may vary across time, especially before and after legislative amendments. LexGenius currently does not incorporate a systematic temporal dimension to assess whether models can understand the time-bound applicability of statutes, the validity period of precedents, or transitional legal provisions. Without such temporal sensitivity tests, models may produce outdated or legally invalid answers when facing evolving legal frameworks, with no mechanism to detect these errors.

I Future Work

While LexGenius establishes a structured evaluation framework for Chinese legal general intelligence, it has yet to fully capture real-world complexity. Therefore, our future work focuses on:

Incorporating Multimodal Tasks to Enhance Realistic Evidence Modeling. The current version of LexGenius relies solely on text and does not include multimodal information common in real legal cases, such as scanned contracts, courtroom audio, or surveillance stills. The absence of such inputs limits the evaluation of model capabilities in visual perception, auditory comprehension, and cross-modal reasoning, essential for evidence review, fact reconstruction, and interpretation of visual-legal content. In the future, we plan to embed images, audio, and other modalities into tasks to assess reasoning capabilities based on heterogeneous, multi-source information, thus aligning evaluation more closely with practical judicial needs.

Expanding Linguistic and Jurisdictional Coverage to Improve Cross-Cultural Generalization. The current dataset is grounded in Chinese texts and the law system of Mainland China, exhibiting limitations in language and legal tradition. This restricts evaluation applicability. Future versions will incorporate texts from Hong Kong, Macau, and Taiwan, as well as English statutes and case law from common law systems. We aim to construct bilingual QA pairs, translation tasks, and comparative analyses to evaluate models' capabilities in understanding, aligning, and adapting across legal and linguistic contexts. This expansion contributes to benchmarking models for global legal services.

Introducing Dynamic Testing of Legal Temporality and Time Sensitivity. Legal applicability is highly time-dependent. Legal amendments can lead to different rulings, and precedents often carry specific periods of validity and applicability. Currently, LexGenius lacks a systematic temporal dimension, making it difficult to evaluate whether a model can identify the applicable time windows of statutes, conditions for transitional provisions, or conflicts between old and new laws. Future versions will include temporally structured legal tasks that require models to make dynamic judgments under varying timeframes, enhancing their understanding and adaptability to evolving legal systems.