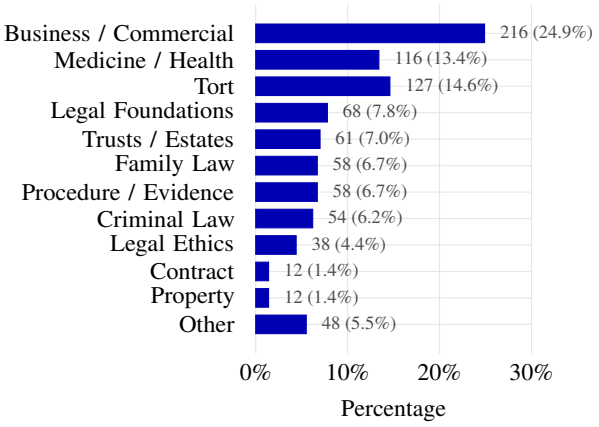
LEGALSCOPE: Measuring Exam-to-Case Transfer in LLM Legal Reasoning

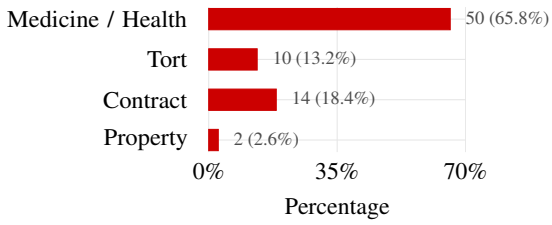
Anonymous ACL submission

Abstract

Public legal exams are widely used to evaluate legal reasoning in large language models, yet it remains unclear whether exam performance reflects the ability to reason over real-case facts. We introduce LEGALSCOPE, a dual-track benchmark that pairs 868 open-ended legal-exam questions with 76 lawyer-reviewed issue-stance prompts derived from de-identified Chinese civil judgments. LEGALSCOPE evaluates models under both reference-answer scoring and case-based rubric scoring, enabling direct analysis of exam-to-case transfer. Across 20 model groups, public-exam scores correlate with real-case scores but do not fully predict them, and real-case prompts expose a recurring weakness in extracting operative legal and factual constraints. We further show that automated evaluation aligns strongly with human review on exam answers but less reliably on real-case legal analysis, highlighting the need for expert-grounded case evaluation.



(a) Bar exam tasks, $n = 868$



(b) Real cases, $n = 76$

Figure 1: Benchmark breakdown of LEGALSCOPE.

1 Introduction

LLMs are increasingly used and evaluated on legal tasks, from document review and question answering to bar-exam style problems (Shi et al., 2026; Pires et al., 2026; Chen et al., 2026; Lee et al., 2026). Standardized benchmarks such as LexGLUE (Chalkidis et al., 2022), LegalBench (Guha et al., 2023), and LawBench (Fei et al., 2024) track this progress at scale, and recent studies show that strong general-purpose LLMs reach high accuracy on professional exams and statutory reasoning tasks (Tafjord et al., 2021; Han et al., 2024; Yang et al., 2018; Zhu et al., 2024; Wu et al., 2024; Trivedi et al., 2022). As a result, public legal exams have become a default proxy for evaluating legal reasoning ability.

However, strong performance on public legal benchmarks does not necessarily indicate strong performance on real-case legal analysis. Public exams are large-scale, standardized, and accompanied by reference answers, whereas real legal practice involves long fact patterns, incomplete information, contested interpretations, and competing lines of argument. A useful legal response must not only recall relevant authorities, but also identify the operative conditions embedded in those authorities, connect those conditions to the facts under a controlled record, and construct a defensible argument under a specified stance. Existing benchmarks rarely pair public-exam evaluation with expert-reviewed reasoning over de-identified real cases, leaving the exam-to-case transfer relationship empirically uncharacterized (Guha et al., 2023; Koreeda and Manning, 2021; Hendrycks

et al., 2021; Han et al., 2025a; Wang et al., 2023; Wilson et al., 2016; Zheng et al., 2021; Zimmeck et al., 2019; Ravichander et al., 2019; Holzenberger and Van Durme, 2021; Lippi et al., 2019; Zheng et al., 2025).

We introduce LEGALSCOPE, a dual-track benchmark designed to study this gap, with its source composition, jurisdiction coverage, and legal-domain distribution summarized in Figure 1. The first track is a public-exam split of 868 open-ended legal-exam instances drawn from the United States, China, the United Kingdom, and Australia. The second track is a lawyer-reviewed real-case split of 76 issue-stance prompts derived from 15 de-identified Chinese civil judgments and covering 38 legal issues, with each issue typically paired into supporting and opposing prompts so that models must construct statute-grounded arguments under a specified stance rather than imitate the observed judgment outcome. The exam track is scored by reference-answer matching, while the real-case track is scored under a calibrated three-dimensional rubric covering citation relevance, constraint extraction, and argument validity, with human validation on overlapping subsets.

Across 20 model groups, public-exam scores correlate with Chinese real-case scores at the model level (Pearson $r = 0.835$), but rankings, reasoning-mode gains, and dimension-level strengths do not transfer uniformly across the two settings. Real cases expose a bottleneck in constraint extraction rather than legal fluency: models score 73.3 on argument validity but only 54.8 on constraint extraction (out of 100), writing coherent arguments yet failing to recover the operative conditions and factual boundaries that control them. Automated evaluation aligns strongly with human review on exam answers (answer-level Pearson $r = 0.925$) but weakens substantially on real-case analysis ($r = 0.450$), indicating that scalable model-judge protocols do not yet match expert legal judgment in case-based settings.

Our contributions are fourfold.

- We introduce LEGALSCOPE, a dual-track legal reasoning benchmark that places scalable public legal-exam evaluation and lawyer-reviewed real-case analysis within a single research framework.
- We provide a paired issue-stance protocol for de-identified Chinese civil judgments, in which the same factual background is evaluated under both a supporting and an opposing stance.
- We design two evaluation mechanisms tailored for LEGALSCOPE: a reference-aware scorer for public exam-style questions and a calibrated three-dimensional rubric for real-case analysis. Both scoring protocols are validated against lawyer review on overlapping subsets.
- We empirically show that exam-to-case transfer is only partial, and identify constraint extraction as the dominant failure mode in real-case settings. We show that model-judge evaluation can diverge substantially from expert review when reasoning depends on case-specific facts, stance requirements, and legal constraints.

2 Related Work

Legal Benchmark Datasets. Legal NLP benchmarks span judgment prediction, retrieval, question answering, entailment, contract review, and reasoning. Early judgment-prediction work on the European Court of Human Rights and CAIL demonstrated the value of large-scale legal text modeling (Aletras et al., 2016; Xiao et al., 2018). Subsequent resources extended evaluation to legal holdings, contracts, unfair clauses, entailment, merger agreements, English legal language understanding, and expert-designed reasoning tasks through CaseHOLD, CUAD, ContractNLI, CLAUDETTE, MAUD, LexGLUE, and LegalBench (Hendrycks et al., 2021; Koreeda and Manning, 2021; Lippi et al., 2019; Wang et al., 2023; Chalkidis et al., 2022; Guha et al., 2023). Chinese resources such as JEC-QA, LeCaRD, LawBench, and LAiW cover Chinese legal reasoning and retrieval (Li et al., 2023; Fei et al., 2024; Dai et al., 2025), while CourtReasoner introduces expert-annotated judicial reasoning over U.S. court opinions (Han et al., 2025b). These benchmarks substantially advance legal NLP evaluation, but they rarely pair scalable public-exam evaluation with expert-reviewed reasoning over real cases. Although real-case benchmarks are closer to professional legal reasoning, they are more expensive to annotate and harder to evaluate automatically. As a result, the relationship between bar exam performance and real-case performance remains empirically undercharacterized, especially when legal reasoning depends on case-specific facts, procedural constraints, and professional judgment. Thus, we design LEGALSCOPE to fill this gap by placing public legal-exam tasks and lawyer-reviewed real-case analysis within a unified evaluation framework.

Benchmark	Public exam	Real cases	Paired stance	Expert review	Exam-to-case transfer
LexGLUE	-	△	-	△	-
CUAD	-	-	-	△	-
LegalBench	✓	-	-	△	-
LawBench	✓	△	-	△	-
CourtReasoner	-	✓	✓	✓	-
Bar Exam	✓	-	-	△	-
LEGALSCOPE	✓	✓	✓	✓	✓

Table 1: Comparison of LEGALSCOPE with representative legal benchmarks. This table summarizes task coverage, not benchmark quality. ✓ denotes full coverage, △ denotes partial coverage, and - denotes no coverage.

Legal Exams as Proxies for Legal Reasoning. Public legal exams are attractive evaluation data because they are large-scale, standardized, and accompanied by official or semi-official reference answers. GPT-4's bar-exam performance demonstrated that strong models can perform well on standardized legal tests (Katz et al., 2024), and follow-up work studies statutory reasoning, tax reasoning, and the interpretation of bar-exam scores (Blair-Stanek et al., 2023; Bommarito and Katz, 2022; Martínez, 2024). However, exam materials may appear in pretraining data (Carlini et al., 2021; Deng et al., 2024; Xu et al., 2024), so exam scores should be treated as comparative signals rather than clean generalization tests. In LEGALSCOPE, we use public exams as a scalable proxy whose limits can be tested against expert-reviewed real-case analysis.

Evaluating Open-Ended Legal Reasoning. Open-ended legal analysis cannot be evaluated by exact match. LLM-as-judge methods offer a scalable alternative, but model-based evaluators are sensitive to position bias, verbosity bias, self-preference, and rubric wording (Liu et al., 2023; Kim et al., 2024). These risks are sharper in legal settings, where fluent prose can still cite the wrong authority, omit a dispositive condition, or invent facts outside the record, and where unsupported authorities and hallucinated citations are a documented failure mode (Dahl et al., 2024; Magesh et al., 2024). In LEGALSCOPE, we therefore pair model-based scoring with lawyer review on overlapping subsets, calibrating our three-dimensional rubric against expert comments so we can study not only whether exam performance transfers to real-case analysis, but also whether automated legal evaluation agrees with expert legal judgment.

3 Benchmark Design

LEGALSCOPE is organized into two complementary tracks designed to study how legal reasoning performance changes across evaluation settings. The public-exam track is a scalable, reference-aware split that mirrors how prior work has measured legal ability on standardized exams. The real-case track is a lawyer-reviewed, closed-book split that asks models to construct statute-grounded arguments over de-identified Chinese civil judgments. Figure 1 summarizes dataset composition, and Figure 2 illustrates the construction pipeline shared by both tracks.

3.1 Public-Exam Split

An instance in the public-exam split is one open-ended legal-exam question paired with a reference answer. Each instance carries the question text, jurisdiction, legal category, reference answer, and a set of review constraints, such as restrictions on which authorities may be cited or which facts may be assumed. A model answer is scored for substantive consistency with the reference rather than surface similarity, with the evaluator inspecting legal issues, governing rules, fact application, exceptions, and the final conclusion. The split contains 868 instances drawn from the United States (604), China (100), the United Kingdom (86), and Australia (78). Sources span U.S. state bar exams along with NCBE and UBE materials, U.K. and Australian bar-equivalent exams, and Chinese legal-examination and training corpora; the full source breakdown appears in Appendix B.2.

3.2 Real-Case Split

The real-case materials are concluded Chinese civil judgments contributed by a practicing Chinese lawyer specializing in medical-damage disputes, who provided cases with direct professional familiarity and reviewed them after de-identification. Each judgment is decomposed into the discrete legal issues addressed in its reasoning, producing 38 issues over 15 judgments. Every issue is rewritten as a closed-book factual background and, where applicable, paired into a supporting prompt and an opposing prompt, yielding 76 issue-stance prompts in total. For each prompt, the model is required to produce two Chinese paragraphs: one summarizing the legally relevant facts and disputed issue, and one applying the relevant authorities to explain the reasoning path under the assigned stance.

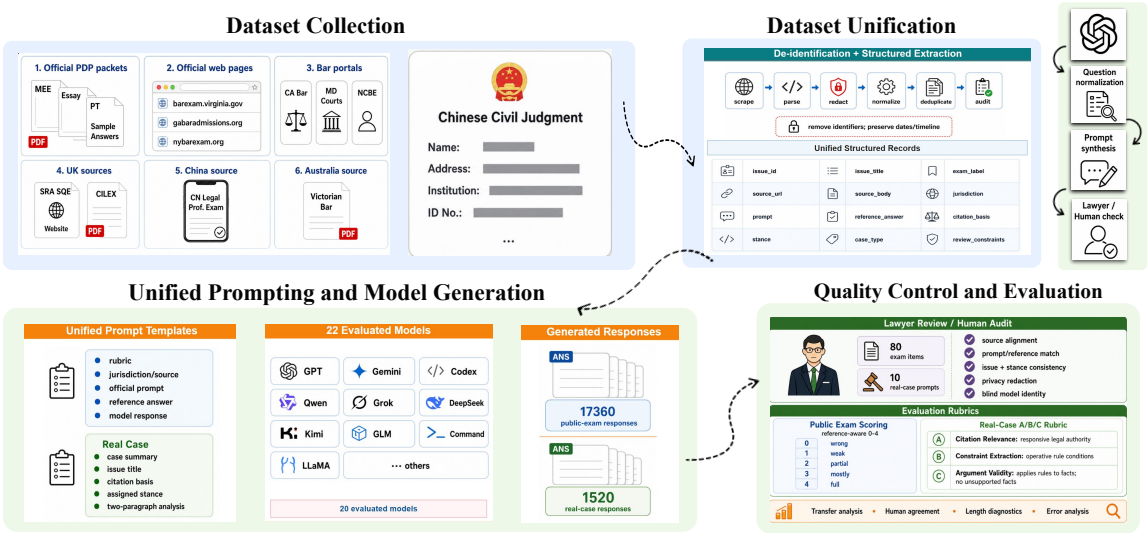


Figure 2: Benchmark construction pipeline. The pipeline first collect raw data from multiple sources, then conduct de-identification, structured extraction. QA pairs are then generated using carefully curated prompts. To ensure quality, human verification and annotation is implemented at all key stages for data collection and evaluation.

Closed-book constraints prohibit external search, undisclosed facts, real identities, and placeholder citations; the prompt also hides the authorities actually cited by the court so that the model must select legally responsive authorities on its own. During de-identification, we remove or mask names, institutions, addresses, identification numbers, contact information, monetary identifiers, and original file paths, and we rewrite the source material as a curated background summary before issue extraction.

3.3 Paired Stance Design

A paired stance consists of two prompts built from the same factual background and legal issue, where one prompt asks the model to support the court's treatment of the issue and the other asks it to oppose that treatment. Judgment-prediction systems can learn direct, sometimes legally irrelevant associations between facts and outcomes (Santosh et al., 2022; Medvedeva and Mcbride, 2023; Chen et al., 2023); paired stance prompts disrupt this shortcut by requiring the model to construct arguments on both sides of the same issue. Supporting prompts reward the ability to assemble the facts and authorities that justify the court's result, while opposing prompts reward the ability to identify gaps, missing conditions, or alternative interpretations. We do not assume that the two stances are equally difficult; the design lets us observe whether a model understands the issue structure rather than only the expected outcome.

4 Evaluation Protocol

LEGALSCOPE uses two scoring protocols matched to the structure of each track: reference-answer matching for the public-exam split, and a calibrated three-dimensional rubric for the real-case split. The following subsections define each scoring scale, the rubric and its calibration path, and the automatic evaluator used throughout the paper.

4.1 Public-Exam Scoring

The public-exam split is scored against the reference answer on an integer scale from 0 to 4. A score of 4 means that the answer covers the decisive legal points and reaches a compatible conclusion; 3 means that the core direction is correct but some elements are missing; 2 means that the answer contains relevant legal content but incomplete application; 1 means only generic or weakly related legal statements; and 0 means that the answer is wrong, irrelevant, or not evaluable. For comparison across tracks, all scores are linearly mapped to a 0–100 scale by multiplying by 25.

4.2 Real-Case Scoring Rubric

Real cases have no single reference answer, so we score them on three dimensions that together summarize whether a response is legally controlled. Citation relevance asks whether the legal authority selected by the answer responds to the issue; constraint extraction asks whether the answer recovers the operative rule conditions and factual bound-

Dimension	Description
Citation Relevance	Does the cited authority address the issue, and is the proposition it supports made explicit?
Constraint Extraction	Does the answer recover the operative rule conditions and factual boundaries?
Argument Validity	Does the conclusion follow from the cited law and facts under the assigned stance?

Table 2: Real-case scoring rubric. Each model answer receives separate integer scores from 0 to 4 on the three dimensions. Detailed high-score and low-score behaviors are given in Appendix A.5.

Rubric stage	Full avg	3/4 share	Human overlap
Initial Checklist	77.3	79.8%	76.8 / 75.5
Strict Regrade	58.0	40.0%	56.3 / 75.5
Calibrated Real-Case Scores	65.3	54.5%	66.8 / 75.5

Table 3: Three-stage calibration of the real-case rubric. The human-overlap column reports automatic mean / human mean on the 200 overlapping real-case answers.

aries; argument validity asks whether the conclusion follows from the cited authorities and facts under the assigned stance. Each dimension receives an integer score from 0 to 4 on the same scale as the public-exam track, and the real-case score is the arithmetic mean of citation relevance, constraint extraction, and argument validity mapped to the 0–100 scale. Table 2 summarizes the three dimensions; complete high-score and low-score behavioral descriptions are given in Appendix A.5.

4.3 Rubric Calibration

We calibrated the real-case rubric in three stages, summarized in Table 3. An initial checklist rubric was easy to apply but produced inflated scores, with 79.8% of dimension scores at 3 or 4 and strong models approaching ceiling performance. A strict regrading rubric corrected the ceiling but over-penalized ordinary omissions, pushing automatic scores far below human scores on the overlapping subset. The final calibrated rubric used throughout the paper preserves strict caps for non-responsive, unreadable, or stance-reversed answers, while allowing limited partial credit for answers that contain usable legal or factual analysis despite a missing authority or incomplete constraint extraction; full cap rules are given in Appendix A.5.

4.4 Automatic Evaluator

We implement the scoring protocols with Codex 5.5 High as the automatic evaluator; the evaluator does not conduct independent legal research. For each scoring instance, the evaluator receives the prompt, the scoring criteria, and a single anonymized model answer, and returns either a 0–4 reference-matching score on the public-exam track or separate 0–4 scores on citation relevance, constraint extraction, and argument validity on the real-case track; the complete field list per track is given in Appendix A.2. The evaluator never sees the model identity, prior scores, or prior review notes. This allows for reduction of anchoring across answers and the generator-evaluator bias that can arise when an LLM judge evaluates text produced by a related model family.

5 Experiments and Results

5.1 Experiment Setup

Models. We evaluate 20 model groups across both tracks, spanning proprietary and open-source families and covering standard and reasoning-mode variants where available. The full model list is provided in Appendix A.1.

Generation. Each model answers the prompts under the task-level constraints described in §3 (closed-book real-case constraints, open-ended exam answering) without modification.

Aggregation and Normalization. Public-exam scores are computed over 868 prompts per model. Real-case scores are computed over 76 prompts per model, with citation relevance, constraint extraction, and argument validity averaged into a real-case mean. All scores are mapped to a 0–100 scale by multiplying the raw 0–4 values by 25 so that the two tracks are directly comparable.

Statistical Reporting. For exam-to-case transfer, we report Pearson and Spearman correlations at the model level, using the 20 model groups as observations. Specific settings on bootstrap confidence intervals are reported in Appendix B.6.

5.2 Main Results

> **RQ1:** How do current LLMs perform across the two tracks of LEGALSCOPE, and how well does exam performance transfer to real cases?

The public-exam mean is 72.0 and the real-case mean is 65.3 across the 20 model groups. The real-case rubric reveals a clear dimension gap: argument validity is the highest at 73.3, citation relevance is moderate at 67.8, and constraint extraction

Model	Public Exam		Real Case					Overall	
	Auto	Human	Citation	Constraint	Argument	Mean	Human	Auto	Human
GPT-5.4 xhigh	87.9	85.0	84.5	60.5	94.4	79.8	80.8	83.9	82.9
Gemini 3 Flash	84.6	83.1	81.2	52.0	74.3	69.2	86.7	76.9	84.9
GPT-5.4 Mini	85.7	75.6	87.8	61.8	91.4	80.4	80.0	83.1	77.8
Gemini 3 Flash Low	84.8	80.6	81.2	50.3	74.3	68.6	84.2	76.7	82.4
GPT-5.4 medium	79.6	75.9	78.9	62.5	90.8	77.4	83.3	78.5	79.6
Kimi K2 Thinking	76.5	69.4	80.3	53.9	75.0	69.7	78.3	73.1	73.9
DeepSeek R1	75.7	71.6	77.3	54.6	75.7	69.2	75.0	72.5	73.3
DeepSeek V3.2	70.0	71.9	72.4	53.9	76.6	67.7	79.2	68.9	75.6
Grok 4.1 Fast	74.2	74.1	57.2	56.2	67.8	60.4	75.8	67.3	75.0
GLM-5	68.8	66.2	74.7	58.2	76.6	69.8	76.7	69.3	71.5
Qwen3 Max Thinking	68.7	62.8	81.2	55.9	77.6	71.6	78.3	70.2	70.6
Deep Cogito 671B	67.5	66.9	65.5	54.6	69.7	63.3	80.8	65.4	73.9
Gemini 2.5 Flash	71.8	62.5	65.5	58.9	74.7	66.3	76.7	69.1	69.6
Grok 4 Fast	74.3	71.9	54.3	54.3	66.4	58.3	72.5	66.3	72.2
Qwen3 Max	69.5	60.9	78.3	57.6	80.9	72.3	73.3	70.9	67.1
Grok 4.20	67.9	67.5	59.9	59.2	72.7	63.9	75.0	65.9	71.3
Qwen3 235B A22B	66.3	62.2	74.7	55.3	75.3	68.4	73.3	67.4	67.8
Command A	62.3	59.7	50.0	52.6	64.8	55.8	71.7	59.1	65.7
Command R+	58.1	56.9	36.2	46.4	53.6	45.4	65.8	51.8	61.4
LLaMA 3.1 8B Instruct	45.4	42.2	14.8	37.2	33.9	28.6	42.5	37.0	42.4

Table 4: Model performance on LEGALSCOPE. Public-exam scores use reference-answer matching; real-case scores use the calibrated rubric. The **Human** columns report lawyer-validated means on overlapping subsets. **Overall** splits the headline summary into two means: **Auto** averages the public-exam **Auto** and real-case **Mean**; **Human** averages the public-exam **Human** and real-case **Human**.

is the lowest at 54.8. The distribution of constraint extraction scores is especially concentrated: 64.8% of constraint extraction answers receive a score of 2 and only 0.9% receive 4, whereas argument validity has 49.3% at 3 and 24.3% at 4. Models can often produce a coherent legal argument once a stance is given, but they reliably struggle to recover the conditions and factual boundaries that make the argument legally controlled, a point that §6.1 develops further.

Exam group	N	Avg	Pearson	Spearman
All public exams	868	72.0	0.835	0.661
China	100	59.3	0.868	0.805
English aggregate	768	73.6	0.820	0.636
United States	604	73.5	0.824	0.637
United Kingdom	86	82.1	0.743	0.458
Australia	78	65.2	0.768	0.538

Table 5: Model-level transfer from public legal exams to Chinese real-case analysis. The real-case endpoint uses the calibrated rubric. English aggregate combines U.S., U.K., and Australian samples.

Public-exam scores correlate with Chinese real-case scores at the model level, but the relationship is not deterministic. Table 5 reports model-level correlations between public-exam means and Chinese real-case means: the full public-exam mean reaches Pearson $r = 0.835$ and Spearman $\rho = 0.661$, and the Chinese exam subset gives the strongest signal at $r = 0.868$, $\rho = 0.805$ despite its lower absolute mean. English-language exam subsets correlate less strongly, which is consistent with the idea that language and legal-system proximity affect transfer. Rankings between the two settings also diverge: top-tier exam models do not always retain their order on the real-case track, and several mid-tier exam models reach competitive real-case scores (Table 4; transfer scatterplots are given in Appendix B.5).

6 Analysis

💡 **RQ2:** Where does real-case legal reasoning fail, and how does it vary across models and case types?

We probe the structure behind the main results through four analyses of where and how real-case legal reasoning fails: a dimension-level breakdown (§6.1), within-family and reasoning-mode contrasts (§6.2), source and domain variation (§6.3), and a failure-mechanism taxonomy with a case study (§6.4). §6.5 then turns to whether automated scoring agrees with expert legal review.

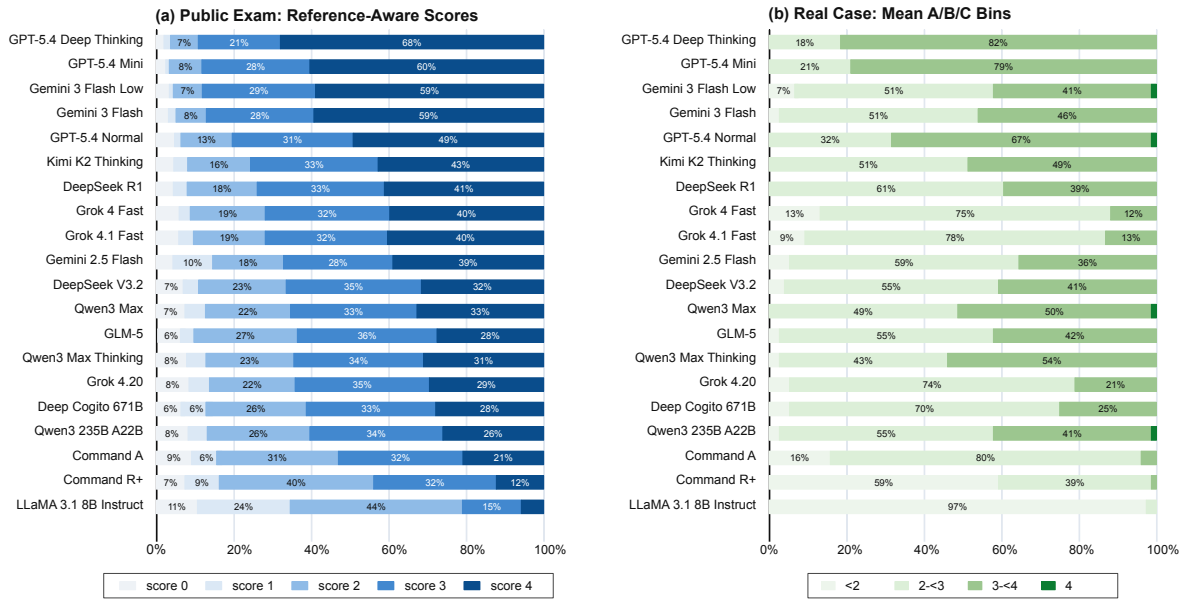


Figure 3: Model stability and score distribution. Public-exam scores show the distribution of reference-aware score bands, while real-case scores show mean A/B/C bins under the calibrated rubric.

Category	N	Citation	Constraint	Argument	Avg
Tort	60	66.7	53.4	72.0	64.1
Contract	10	71.1	61.9	79.3	70.8
Property	6	73.3	56.7	76.5	68.8

Table 6: Chinese real-case performance by category. N counts issue-stance prompts; each prompt is answered by 20 model groups.

Comparison	Exam Δ	Citation Δ	Constr. Δ	Avg Δ
GPT-5.4 xhigh - GPT-5.4 Mini	+2.2	-3.3	-1.3	-0.6
GPT-5.4 xhigh - GPT-5.4 medium	+8.3	+5.6	-2.0	+2.4
Gemini 3 Flash - Low	-0.2	+0.0	+1.7	+0.6
Qwen3 Max Thinking - Non-Thinking	-0.8	+2.9	-1.7	-0.7
DeepSeek R1 - DeepSeek V3.2	+5.7	+4.9	+0.7	+1.5
Grok 4.1 Fast - Grok 4 Fast	-0.1	+2.9	+1.9	+2.1

Table 7: Within-family and reasoning-mode comparisons. Positive values mean the first model in the comparison scores higher. **Constr.** abbreviates **Constraint**.

6.1 Dimension-Level Analysis

The dimension gap reported in §5.2 is the dominant real-case finding: models produce fluent legal arguments once given a direction but fail to recover the operative conditions that make a rule applicable, such as expert appraisal, causation, liability share, payment milestone, or evidence boundary. Public-exam scoring can miss this failure mode because reference-aware scoring rewards substantive coverage of the official answer without requiring explicit recovery of fact-bounded conditions. Table 6 shows the bottleneck systematically across case categories: contract prompts, which expose concrete transactional conditions, score higher on constraint extraction than medical-damage tort prompts, where models more often miss controlling conditions like appraisal opinions, causal contribution, and liability ratios.

6.2 Reasoning-Mode and Family Analysis

Extending the dimension-level analysis with a within-family view, we ask whether reasoning-oriented or larger model variants improve real-case constraint extraction. Table 7 compares closely related model variants across six families. Reasoning-oriented variants typically improve public-exam performance, but the gain does not transfer uniformly to real-case analysis. GPT-5.4 xhigh scores $+2.2$ points above GPT-5.4 Mini on the public exam but -0.6 points on the Chinese real-case mean; relative to GPT-5.4 medium it gains $+8.3$ on the public exam yet only $+2.4$ on the real-case mean, and is -2.0 on constraint extraction. Outside the GPT family, Qwen3 Max Thinking gains $+2.9$ on citation relevance but loses on constraint extraction, argument validity, and the real-case mean, and DeepSeek R1 gains $+5.7$ on the public exam but only $+1.5$ on the real-case mean. These within-family contrasts indicate that reasoning mode is not a single uniform advantage: it can sharpen long-form argument construction without solving the constraint-extraction problem that defines the real-case bottleneck.

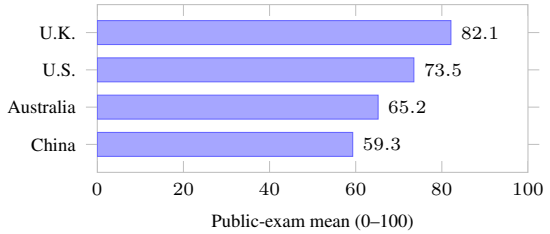


Figure 4: Public-exam means by country jurisdiction across 20 model groups. U.S. subsource and topic-category breakdowns are reported in Appendix B.2.

6.3 Source and Domain Analysis

Extending the dimension-level analysis with a cross-source view, we look at how public-exam performance varies across country jurisdictions, U.S. subsources, and legal categories. These differences should be read as source and format effects rather than intrinsic jurisdictional difficulty. Figure 4 summarizes country-level means; the full U.S. source and topic breakdowns are reported in Appendix B.2. U.K. samples score highest at 82.1, followed by U.S. at 73.5, Australia at 65.2, and China at 59.3; within the U.S. portion, New York/NCBE MEE and Maryland UBE samples lead while California fictional State of Columbia items, which require following prompt-specific fictional authority, score lowest. These variations modulate but do not redefine the partial-transfer and constraint-bottleneck conclusions above.

6.4 Error Analysis

Extending the dimension-level analysis with a failure-mechanism view, the real-case scoring notes reveal five recurring error patterns: citation mismatch, condition omission, fact extrapolation, stance drift, and conclusion jump. The shared problem is not that models lack legal vocabulary; many answers cite plausible rules and write fluent legal prose. Rather, the rules are often too broad, the conditions are incomplete, or the answer adds facts the prompt did not provide, so the conclusion appears persuasive while resting on a missing condition or unsupported authority. We include the per-mechanism behavior and dimension impact, representative anonymized error cards, and a worked case study on RV038 in Appendix C.1.

6.5 Human Validation

RQ3: Can model-judge evaluation reliably approximate expert legal review?

Validation set	Model	Human	Answer-level
Public exams	70.7	68.3	$r = .925, \rho = .928$
Real cases	66.8	75.5	$r = .450, \rho = .384$

Table 8: Automatic scoring versus human validation. Model-level correlations are $r = .992, \rho = .986$ for public exams and $r = .874, \rho = .669$ for real cases.

Human validation calibrates the automatic scores. The public-exam validation subset contains 80 exam items and 20 model groups, for 1,600 answers; the real-case validation subset contains 10 issue-stance prompts and 20 model groups, for 200 answers. The two subsets differ in reviewer background: the real-case validation was scored by the cooperating lawyer who had direct professional familiarity with the underlying civil disputes and reviewed the de-identified materials, whereas the public-exam validation was scored by human reviewers without the same specialized case knowledge. Table 8 reports agreement. Reference-answer exam scoring aligns strongly with human review; real-case scoring is harder, with the model-level correlation remaining useful but answer-level agreement dropping because experts tolerate some legally plausible reasoning that the rubric penalizes, while the rubric more strictly marks missing conditions and unsupported citations. On the real-case overlap subset, the automatic mean is 66.8 against a human mean of 75.5; we do not interpret this 8.7-point gap as a simple error in either direction but as the remaining distance between scalable rubric-based evaluation and expert legal judgment.

7 Conclusion

We introduced LEGALSCOPE, a dual-track legal-reasoning benchmark that pairs scalable public-exam questions with lawyer-reviewed real-case issue-stance prompts derived from de-identified Chinese civil judgments. The benchmark reveals that exam-style success can coexist with failures to preserve factual boundaries, recover operative conditions, and reason under a constrained case record. The implication is not that public legal exams should be abandoned but that they should be paired with expert-grounded real-case diagnostics. LEGALSCOPE offers one way to measure this gap and to distinguish broad legal knowledge from the ability to apply that knowledge to case-specific facts, constraints, and advocacy positions.

Limitations

We see several directions for extending LEGALSCOPE. First, the real-case split currently focuses on Chinese civil judgments, providing depth under one legal system; extending the protocol to other jurisdictions, languages, and doctrinal areas such as criminal, administrative, and cross-border law is a natural next step. Second, the real-case split contains 76 issue-stance prompts derived from 15 judgments, providing a deep test under medical-damage liability and adjacent civil disputes; scaling the protocol to additional case categories with collaborating practitioners would broaden the doctrinal coverage of the diagnostic. Third, public-exam materials may appear in pretraining data, so we treat exam scores as comparative signals rather than clean generalization tests; future releases can introduce contamination-controlled exam pools or held-out items collected after a freeze date. Fourth, human validation, while large enough to calibrate scores and model rankings, does not support full inter-annotator agreement or expert-disagreement adjudication; subsequent work can add independent double annotation and finer error labels for citation hallucination, condition omission, fact supplementation, stance reversal, and unsupported conclusion jumps (Dahl et al., 2024; Magesh et al., 2024).

Ethics and Data Governance

The real-case prompts are derived from de-identified Chinese civil judgments, with names, institutions, addresses, identification numbers, contact information, monetary identifiers, and original file paths removed or masked before issue extraction; the benchmark does not ask models to predict real litigants or recover identities, and prohibits external search and undisclosed factual supplementation. LEGALSCOPE is a research benchmark and is not a legal-advice system: a high score means a model performed well under a controlled evaluation protocol, not that it is competent to advise a real client, decide a case, or replace legal professionals. The core risk is that strong legal-exam performance does not guarantee accurate or controlled legal reasoning on real cases; models can produce fluent and legally plausible analysis while citing the wrong authority, omitting a dispositive condition, or introducing facts outside the record, which is exactly the failure pattern LEGALSCOPE is designed to surface. We mitigate these risks through closed-book prompts, systematic de-identification, lawyer review of the real-case materials, calibrated three-dimensional scoring, and diagnostic rather than predictive framing of the benchmark results.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2:e93.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? *Preprint*, arXiv:2302.06100.

Michael James Bommarito and Daniel Martin Katz. 2022. Gpt takes the bar exam. *ArXiv*, abs/2212.14402.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. *Preprint*, arXiv:2012.07805.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. *Preprint*, arXiv:2110.00976.

Haotian Chen, Lingwei Zhang, Yiran Liu, Fanchao Chen, and Yang Yu. 2023. Knowledge is power: Understanding causality makes legal judgment prediction models more generalizable and robust. *Preprint*, arXiv:2211.03046.

Sijia Chen, Hang Yin, and Shunfan Zhou. 2026. Legalcitebench: Evaluating citation reliability in legal language models. *Preprint*, arXiv:2605.10186.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. Laiw: A chinese legal large language models benchmark. In *Proceedings of the 31st International conference on computational linguistics*, pages 10738–10766.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages

8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

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.

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.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Simeng Han, Stephen Xia, Grant Zhang, Howard Dai, Chen Liu, Lichang Chen, Hoang Huy Nguyen, Hongyuan Mei, Jiayuan Mao, and R Thomas McCoy. 2025a. Creativity or brute force? using brainteasers as a window into the problem-solving abilities of large language models. *arXiv preprint arXiv:2505.10844*.

Sophia Simeng Han, Yoshiki Takashima, Shannon Zejiang Shen, Chen Liu, Yixin Liu, Roque K Thuo, Sonia Knowlton, Ruzica Piskac, Scott J Shapiro, and Arman Cohan. 2025b. Courtreasoner: Can llm agents reason like judges? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35279–35294.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Nils Holzenberger and Benjamin Van Durme. 2021. Factoring statutory reasoning as language understanding challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2742–2758.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 382.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919.

Jinu Lee, Kyoung-Woon On, Simeng Han, Arman Cohan, and Julia Hockenmaier. 2026. Evaluating legal reasoning traces with legal issue tree rubrics. *Preprint*, arXiv:2512.01020.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023. Lecardv2: A large-scale chinese legal case retrieval dataset. *Preprint*, arXiv:2310.17609.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139.

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

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *Preprint*, arXiv:2405.20362.

Eric Martínez. 2024. Re-evaluating gpt-4's bar exam performance: Re-evaluating gpt-4's bar exam performance. *Artif. Intell. Law*, 33(3):581–604.

Masha Medvedeva and Pauline Mcbride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84, Singapore. Association for Computational Linguistics.

Ramon Pires, Thales Sales Almeida, Celio Larcher Junior, Giovana Bonás, Hugo Abonizio, Marcos Piau, Roseval Malaquias Junior, Thiago Laitz, and Rodrigo Nogueira. 2026. Magis-bench: Evaluating llms on magistrate-level legal tasks. *Preprint*, arXiv:2605.08437.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*.

Tokala Yaswanth Sri Sai Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding legal judgment prediction for european court of human rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138.

Yuzhen Shi, Huanghai Liu, Yiran Hu, Gaojie Song, Xinran Xu, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Di Feng, Wenbo Lv, Weiheng Wu, Kexin Yang, Sen Yang, Wei Wang, Rongyao Shi, Yuanyang Qiu, Yuemeng Qi, Jingwen Zhang, Xiaoyu Sui, Yifan Chen, Yi Zhang, An Yang, Bowen Yu, Dayiheng Liu, Junyang Lin, Weixing Shen, Bing Zhao, Charles L. A. Clarke, and Hu Wei. 2026. Plawbench: A rubric-based benchmark for evaluating llms in real-world legal practice. *Preprint*, arXiv:2601.16669.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Steven Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dmitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. 2023. Maud: An expert-annotated legal nlp dataset for merger agreement understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16369–16382.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. 2024. Cofca: A step-wise counterfactual multi-hop qa benchmark. *Preprint*, arXiv:2402.11924.

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

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. *Preprint*, arXiv:2406.04244.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D Manning, Peter Henderson, and Daniel E Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 169–193.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. FanOutQA: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37, Bangkok, Thailand. Association for Computational Linguistics.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66.

A Evaluation Setup

A.1 Model Groups

The benchmark evaluates the following 20 model groups: GPT-5.4 xhigh, GPT-5.4 Mini, GPT-5.4 medium, Gemini 3 Flash, Gemini 3 Flash Low, Gemini 2.5 Flash, Kimi K2 Thinking, DeepSeek R1, DeepSeek V3.2, Grok 4.1 Fast, Grok 4 Fast, Grok 4.20, GLM-5, Qwen3 Max Thinking, Qwen3 Max, Qwen3 235B A22B, Deep Cogito 671B, Command A, Command R+, and LLaMA 3.1 8B Instruct.

A.2 Real-Case Scoring Packet Fields

Each real-case scoring packet contains exactly one model answer and the following case fields: `Prompt (for AI)`, `Scoring prompt`, `Position`, `Core Issue`, `Citation Basis`, `Cited Supported Proposition`, and `Review constraints`. The scorer does not receive the model name, previous dimension scores, or previous review notes. This packet design is intended to reduce anchoring across model answers and prevent model-family labels from affecting the evaluation.

A.3 Prompt and Scoring Packet Example

Table 9 shows a compressed example of the information visible to the evaluator. The evaluator receives one model answer at a time, together with the case fields needed for scoring. The evaluator does not receive the model name, previous dimension scores, or previous review notes.

A.4 Example of Model Input and Output

This appendix shows representative input-output structures for the two benchmark tracks. The Chinese real-case split uses closed-book issue-stance prompts and a three-dimensional calibrated rubric, while the public bar-exam split uses reference-answer scoring.

A.5 Rubric Calibration Details

Table 10 gives the full high-score and low-score behaviors used when applying the three real-case dimensions. Table 11 then summarizes the three-stage calibration path: the initial checklist rubric was useful for implementation but produced a ceiling effect, the strict regrade corrected the ceiling but over-penalized ordinary omissions, and the calibrated rubric maintains a high standard for score 4 while removing unnecessary global caps. Specifically, the calibrated rubric does not treat ordinary omissions as non-responses, sets citation relevance to 0 when no legal authority is given while allowing constraint extraction and argument validity to retain limited credit when substantive analysis is present, and caps argument validity at 2 rather than forcing it to 0 when constraint extraction is 0.

B More Statistics

B.1 Score Distribution by Dimension

Table 12 reports the score distribution for each real-case dimension. Constraint extraction has the clearest bottleneck: 985 of 1,520 answers receive a score of 2, and only 14 receive a score of 4. This distribution supports the main-text interpretation that the real-case split is not primarily measuring whether models can produce fluent legal prose. It measures whether they can recover the operative rule conditions and factual boundaries that control the legal argument.

B.2 Source and Domain Breakdowns

This appendix reports the full source- and topic-level breakdowns that contextualize the country-level summary in Figure 4. Table 14 shows mean scores across U.S. subsources, including UBE materials, NCBE MEE reprints, and state-level bar exams. Table 15 reports public-exam category means for categories with at least 20 samples.

B.3 Case-Level Difficulty

Table 16 reports case-level difficulty patterns. Means and standard deviations are computed across the 20 model answers for each review ID, after averaging citation relevance, constraint extraction, and argument validity and converting to the 0–100 scale. The lowest-mean cases are concentrated in medical-malpractice and medical-damage disputes, while the highest-disagreement cases often involve appraisal, causation, successor liability, or damages boundaries.

B.4 Human-Model Disagreement

Table 17 lists representative model-level disagreements on the 200 overlapping real-case answers. Rank difference is computed as the calibrated rank minus human rank, so positive values indicate models ranked higher by human review than by the calibrated rubric, and negative values indicate models ranked higher by the calibrated rubric than by human review. The pattern supports the main-text interpretation: the calibrated rubric is stricter about

Packet field	Compressed example from RV038
Issue	Whether the court should permit re-appraisal when the appraisal opinion is alleged to be unfair.
Position	Supports the court judgment.
Core issue	The court did not grant the plaintiff's application for re-appraisal and still used the existing appraisal materials as the basis for judgment.
Citation basis	Article 40 of the Several Provisions of the Supreme People's Court on Evidence in Civil Proceedings.
Cited proposition	Re-appraisal requires statutory conditions such as serious procedural illegality or an appraisal opinion that clearly lacks sufficient basis; dissatisfaction with the conclusion alone is not enough.
Review constraints	Do not add off-topic facts, undisplayed evidence, unrecorded diagnosis or treatment details, external search results, real names, or institutional identities.
Answer visibility	A single model answer is shown without model identity or old scoring information.

Table 9: Compressed example of a real-case scoring packet.

Dimension	High-score behavior	Low-score behavior
Citation relevance	Identifies relevant statutes, judicial interpretations, or legal rules and explains the proposition they support.	Gives no legal authority, cites irrelevant or unsupported authority, or lists authorities without explaining their connection to the dispute.
Constraint extraction	Extracts conditions such as actors, conduct, fault, causation, timing, burden of proof, exceptions, expert opinion, liability share, payment milestones, warranty reserve, acceptance, limitation interruption, and evidence boundaries.	Omits key conditions, ignores fact boundaries, adds unstated facts, confuses the stance, or violates format and privacy constraints.
Argument validity	Maps rule conditions to facts, handles uncertainty, and forms a legally defensible reason under the assigned stance.	Jumps from abstract rules to conclusion, extrapolates facts, breaks the reasoning chain, reverses the stance, or reaches a conclusion incompatible with the cited authority and facts.

Table 10: Detailed high-score and low-score behaviors for each real-case rubric dimension. This expands the summary in Table 2.

unsupported citations, fact supplementation, and missing constraints, while expert review may give more credit to partially useful legal analysis.

B.5 Transfer Scatterplots

Figure 5 visualizes the model-level relationship between public-exam scores and Chinese real-case scores under two scoring settings. Panel (a) uses model-judge scores, while Panel (b) uses human-validation means. These figures complement the correlation statistics reported in the main text.

B.6 Bootstrap Robustness

We estimate bootstrap confidence intervals by resampling the 20 model groups with replacement 10,000 times. Table 18 reports model-level intervals for the exam-to-case transfer correlations.

C Error Analysis

C.1 Qualitative Error Cards

Table 19 maps each of the five real-case failure modes from §6.4 to its typical behavior and the rubric dimensions it most affects. Table 20 then summarizes representative error cards selected from the scoring notes. These examples are anonymized at the case-fact level and are intended to explain error mechanisms, not to provide a complete qualitative audit of any single model.

Case Study: RV038. RV038 illustrates the constraint-extraction-bottleneck mechanism concretely. The issue asks whether a court should permit re-appraisal in a medical-malpractice dispute after an existing appraisal had already addressed fault, causation, and causative potency. Several model answers identified relevant medical-liability rules and produced coherent arguments about appraisal review, yet the main failure was not a lack of legal vocabulary; the weaker answers supplemented the closed-book record with assumed appraisal defects, unstated medical findings, or stronger causal claims than the prompt provided. Under the calibrated rubric, this pattern can receive substantial credit for citation relevance or argument form while still losing points on constraint extraction because the answer does not preserve the factual and evidentiary boundaries of the case. Table 21 gives the structured mechanism breakdown, showing how a response can look legally

Rubric stage	Full avg	3/4 share	Human overlap	Main limitation
Initial checklist	77.3	79.8%	76.8 / 75.5	Too permissive; many fluent answers received high scores even when they missed key constraints.
Strict regrade	58.0	40.0%	56.3 / 75.5	Too harsh; ordinary omissions were sometimes treated like severe non-responses.
Calibrated rubric	65.3	54.5%	66.8 / 75.5	Final setting.

Table 11: Rubric calibration for the Chinese real-case split. The human-overlap column reports automatic mean / human mean on the 200 overlapping real-case answers.

Dimension	0	1	2	3	4
Citation relevance	34	148	383	612	343
Constraint extraction	9	115	985	397	14
Argument validity	8	52	342	749	369

Table 12: score counts by dimension over 1,520 Chinese real-case answers.

Score pattern	Count	Share
3/2/3	270	17.8%
2/2/3	164	10.8%
4/2/3	97	6.4%
4/2/4	97	6.4%
2/2/2	90	5.9%
4/3/4	88	5.8%
3/3/4	84	5.5%
3/3/3	76	5.0%

Table 13: Most frequent score combinations in the Chinese real-case split. Each pattern lists citation relevance, constraint extraction, and argument validity scores in that order.

plausible while failing the real-case requirement that legal reasoning remain tied to the supplied record.

U.S. source / legal-system label	N	Avg
New York / NCBE MEE reprint	102	76.6
Maryland UBE law	75	76.5
Virginia / U.S. bar law	183	73.4
California general U.S. law	100	73.1
Uniform Bar Examination law	57	73.0
Georgia bar exam law	70	69.3
California fictional State of Columbia	17	64.3

Table 14: U.S. public-exam source differences. Means are model-answer averages on the 0–100 scale. Small sources should not be over-interpreted.

Legal category	N	Avg
Foundations of Legal Knowledge	68	84.1
Business Organizations / Agency	126	74.5
Secured Transactions / Commercial Law	67	74.0
Tort	243	73.3
Family	58	72.1
Professional Responsibility / Legal Ethics	38	69.8
Succession / Trusts / Estates	61	69.6
Administrative Law	21	68.6
Civil Procedure	26	67.6
Civil / Civil Procedure	22	60.2
Criminal Procedure	23	56.0
Commercial / Company Law	23	54.8

Table 15: Public-exam performance by legal category. The table includes categories with at least 20 samples.

Rank	Lowest-avg case	Avg/SD	Highest-avg case	Avg/SD	Highest-disagreement case	Avg/SD
1	RV024	47.1/11.6	RV008	78.8/13.8	RV050	71.2/22.7
2	RV053	50.8/13.7	RV006	77.1/16.2	RV061	52.1/20.9
3	RV061	52.1/20.9	RV002	76.7/12.8	RV062	63.8/19.9
4	RV047	54.2/11.3	RV031	75.8/16.6	RV072	63.8/18.7
5	RV023	55.0/15.2	RV033	75.4/13.8	RV045	62.9/18.5
6	RV075	55.0/17.8	RV038	75.4/15.5	RV044	62.9/18.3
7	RV052	56.2/18.2	RV003	74.6/12.8	RV064	61.7/18.3
8	RV021	56.7/10.1	RV028	73.8/13.0	RV056	72.5/18.3
9	RV025	56.7/9.4	RV034	73.8/16.7	RV052	56.2/18.2
10	RV026	56.7/12.8	RV042	73.8/14.3	RV011	62.5/18.0

Table 16: Case-level difficulty and disagreement in the Chinese real-case split. Each mean and standard deviation is computed across 20 model answers for the same review ID.

Model	the calibrated rubric	Human	Gap	Rank diff.
Qwen3 235B A22B	71.7	73.3	-1.7	-10
Deep Cogito 671B	68.3	80.8	-12.5	+8
Gemini 3 Flash Low	70.0	84.2	-14.2	+8
Qwen3 Max	70.8	73.3	-2.5	-7
Gemini 3 Flash	70.8	86.7	-15.8	+6
GPT-5.4 Mini	80.8	80.0	+0.8	-5
Qwen3 Max Thinking	74.2	78.3	-4.2	-5

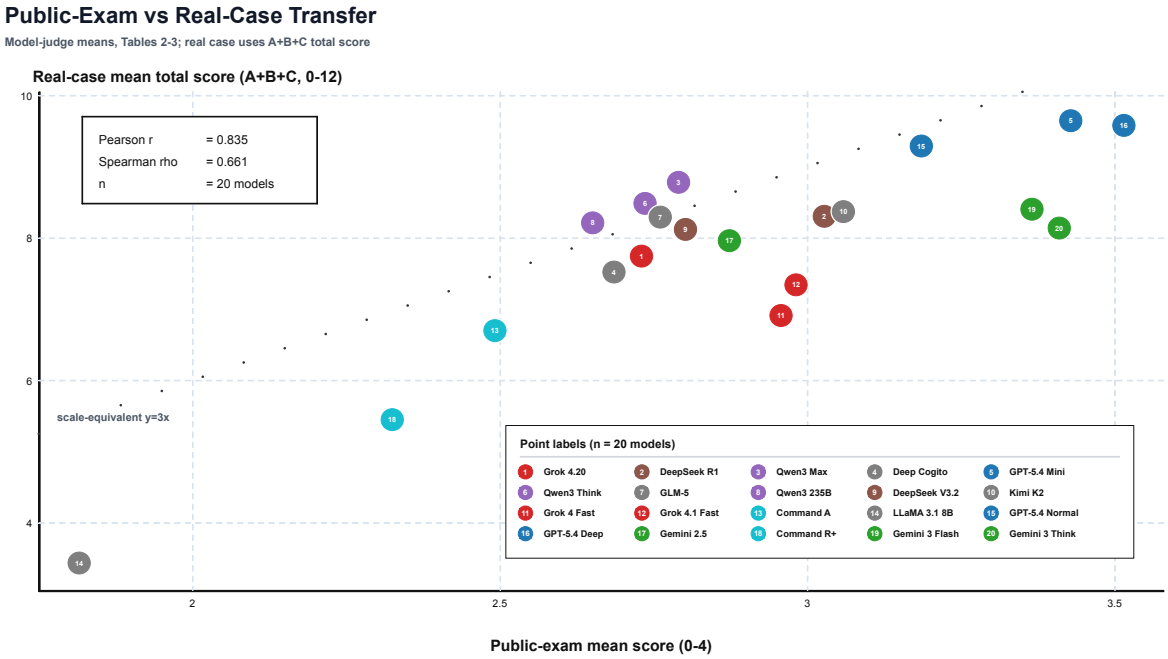
Table 17: Representative model-level disagreements between the calibrated scores and human real-case validation on the overlap subset.

Exam group	Pearson	95% CI	Spearman	95% CI
All exams	0.835	[0.479, 0.938]	0.661	[0.235, 0.896]
China	0.868	[0.688, 0.947]	0.805	[0.488, 0.950]
U.S.	0.824	[0.444, 0.936]	0.637	[0.212, 0.880]
U.K.	0.743	[0.206, 0.899]	0.458	[-0.032, 0.794]
Australia	0.768	[0.416, 0.899]	0.538	[0.030, 0.852]

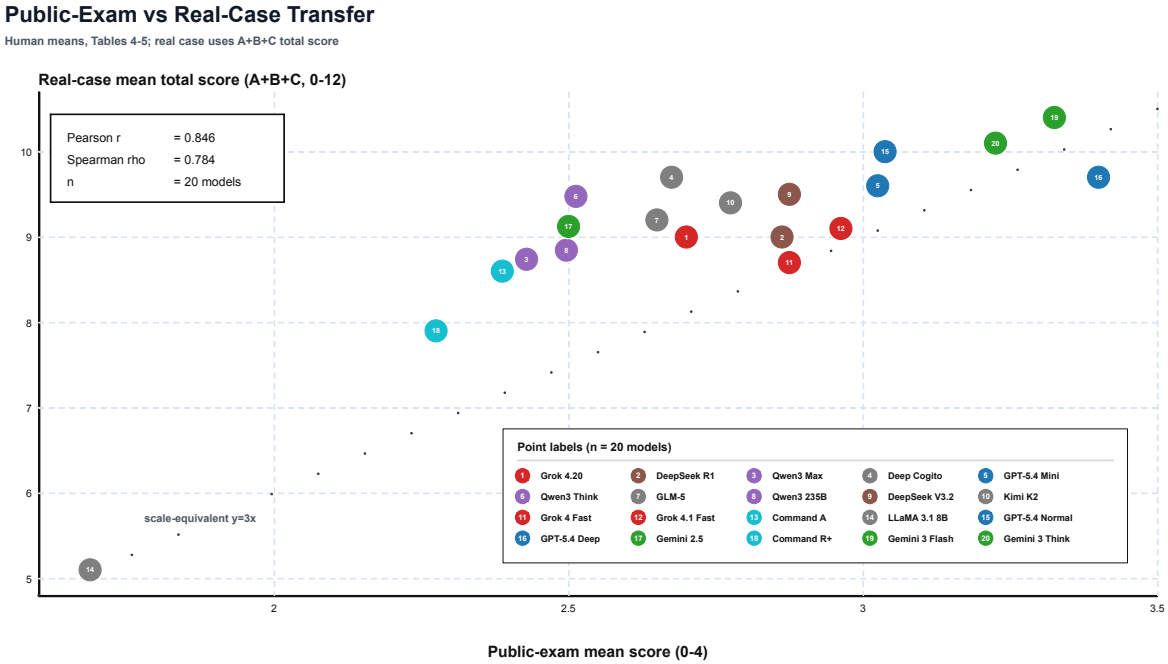
Table 18: Model-level bootstrap confidence intervals for transfer from public-exam scores to Chinese real-case scores.

Mechanism	Typical behavior	Dimensions affected
Citation mismatch	The cited authority is irrelevant, outdated, too broad, or unable to support the proposition.	Citation relevance; often argument validity
Condition omission	The answer omits operative conditions such as actors, fault, causation, timing, burden of proof, exceptions, or liability share.	Constraint extraction; often argument validity
Fact extrapolation	The answer adds expert opinions, attachments, evidence, or performance facts not present in the prompt.	Constraint extraction and argument validity
Stance drift	The answer supports the court when asked to oppose it, or shifts stance during reasoning.	Constraint extraction and argument validity (severe cap)
Conclusion jump	The answer moves from an abstract rule to a conclusion without fact mapping.	Argument validity

Table 19: The five real-case failure modes discussed in §6.4, with typical model behavior and the rubric dimensions most affected.



(a) Model-judge exam-to-case transfer.



(b) Human-scored exam-to-case transfer.

Figure 5: Exam-to-case transfer scatterplots. Each point represents one model group. Panel (a) compares public-exam mean score with Chinese real-case total score under model-judge evaluation. Panel (b) reports the analogous relationship using human-validation means. Panel (b) measures transfer under human scores, not model-vs-human agreement.

Failure mechanism	Review ID	Model	Scores	Short explanation
Citation mismatch	RV001	Command R+	1/1/1	The answer relies on generic or outdated contract authority and adds an unsupported after-sales-service prerequisite, so the cited law does not support the assigned opposition argument.
Condition omission	RV063	Gemini 3 Flash Low	4/1/3	The answer gives a legally plausible damages discussion but does not control causation, successor liability, liability ratio, and damage-item boundaries required by the prompt.
Fact extrapolation	RV038	Qwen3 235B A22B	4/1/3	The answer cites relevant appraisal authority and gives usable re-appraisal reasoning, but it fabricates detailed appraisal defects not present in the closed-book record.
Stance drift	RV001	Command R+	1/1/1	The response begins from the assigned opposition posture but shifts toward a conclusion that does not squarely oppose the court's handling outcome.
Conclusion jump	RV063	Gemini 3 Flash Low	4/1/3	The answer moves from general damages rules to a result without sufficiently mapping the prompt's causation and liability-share constraints.
Format / constraint violation	RV002	Kimi K2 Thinking	4/2/4	The answer identifies the acceptance and payment-milestone structure well, but violates the required Chinese-language format; constraint extraction is reduced without destroying legal validity.
No usable legal basis	RV011	LLaMA 3.1 8B Instruct	0/0/0	The answer gives no usable statutory or judicial-interpretation basis and contains severe garbling and unstable factual phrasing, so the calibrated rubric assigns all-zero scores.

Table 20: Representative qualitative error cards from the Chinese real-case scoring notes, grouped by failure mechanism.

Element	RV038 mechanism
Case issue	Whether re-appraisal should be permitted in a medical-malpractice dispute after an existing appraisal had addressed fault, causation, and causative potency.
Expected constraint	The answer should reason from the displayed appraisal posture and avoid adding undisclosed defects, medical records, or alternative expert findings.
Observed behavior	Some answers cite relevant medical-liability or appraisal rules and give a plausible legal path, but infer extra appraisal deficiencies or stronger causation facts not stated in the prompt.
Rubric consequence	Citation relevance and argument validity can remain moderate or high, but constraint extraction is reduced because the answer crosses the closed factual and evidentiary boundary.

Table 21: Case study of the low-constraint-extraction/high-argument-validity pattern in RV038. The case illustrates the difference between fluent legal argumentation and controlled constraint extraction.