AppealCase: A Dataset and Benchmark for Civil Case Appeal Scenarios

Yuting Huang[1], Meitong Guo[1], Yiquan Wu[1], Ang Li[1], Mengze Li[1]
Xiaozhong Liu[2], Keting Yin[1], Changlong Sun[1], Kun Kuang[1*]

[1]Zhejiang Univeristy, Hangzhou, China
[2]Worcester Polytechnic Institute, Worcester, USA
{yutinghuang, guomeitong, wuyiquan, leeyon, mengzeli, yinkt, kunkuang}@zju.edu.cn
xliu14@wpi.edu, changlong.scl@gmail.com

Abstract

Recent advances in Legal Artificial Intelligence (LegalAI) have focused on single-case judgment analysis while largely overlooking the appellate process. Appeals serve as a vital mechanism for error correction and fair trials, making them crucial in both legal practice and AI research. The appellate scenario presents unique challenges for LLMs, including cross-trial factual dependencies, longer input contexts, and more fine-grained, complex legal reasoning. To address this gap, we introduce AppealCase, a dataset of 10,000 real-world pairs of matched first-instance and second-instance documents across 91 civil categories. AppealCase provides a dedicated annotation scheme along five key dimensions: judgment reversals, reversal reasons, cited legal provisions, claim-level decisions, and whether new information appears in the second instance. Based on these structured annotations, we define five benchmark tasks and evaluate 20 mainstream LLMs. Results show that current models struggle in the appellate setting—on Judgment Reversal Prediction, all models achieve F1 scores below 50%—highlighting the complexity and difficulty of appeal-focused LegalAI. We hope AppealCase fosters future work on appellate case understanding and contributes to more consistent judicial outcomes.

1 Introduction

LegalAI has rapidly advanced from early rule-based systems and logic-based reasoning (Sergot et al. 1986), through statistical learning models (Chalkidis et al. 2020), to the recent emergence of LLMs (Zhou et al. 2024). LLMs have significantly broadened the scope of LegalAI, achieving state-of-the-art results in tasks such as legal consultation Q&A (Büttner and Habernal 2024), and similar case retrieval (Wiratunga et al. 2024). Their strong language understanding and reasoning capabilities have made them the foundation for a new generation of general-purpose legal AI systems.

Despite the recent remarkable progress in LegalAI, most existing research remains focused on first-instance cases. Tasks such as legal judgment prediction (Tong et al. 2024) and court view generation (Li et al. 2024b) are typically framed around single-instance decisions, overlooking appellate proceedings—a structurally essential component of the legal system that enables error correction, clarifies legal standards, and safeguards litigants' rights. Consequently, the reasoning and automation of appeal cases have received insufficient attention in the current LegalAI landscape.

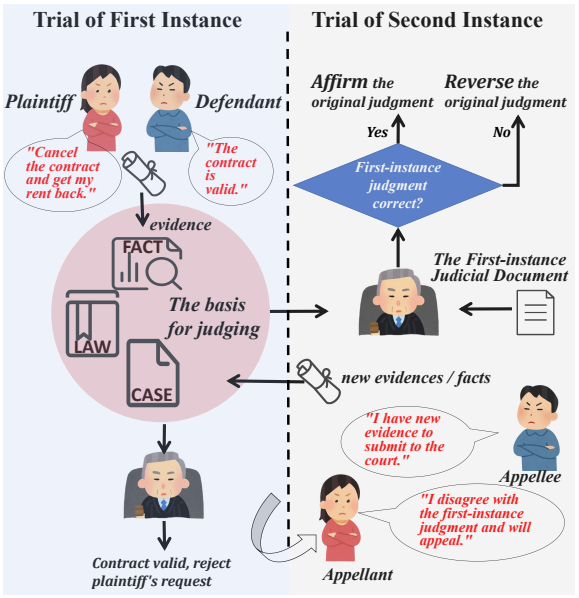

Figure 1: Procedural flow from first-instance trial to second-instance judgment. When a party is dissatisfied, they file an appeal and may submit new evidence if available. The appellate court reviews the previous judgment for errors and decides to either uphold or overturn the decision.

The appellate process is fundamentally distinct from first-instance trials. As shown in Figure 1, in an appeal, the dissatisfied party from the first-instance judgment—now termed the appellant—may challenge the decision, while the opposing party becomes the appellee. The appellate court must not only review the factual findings of the lower court, but also consider new evidence, re-evaluate legal interpretations, and assess the consistency of the original judgment with statutory standards (Merryman and Pérez-Perdomo 2018). This multi-layered review introduces significantly greater demands in reasoning, legal coherence, and information integration. From a modeling perspec-

*Corresponding author.

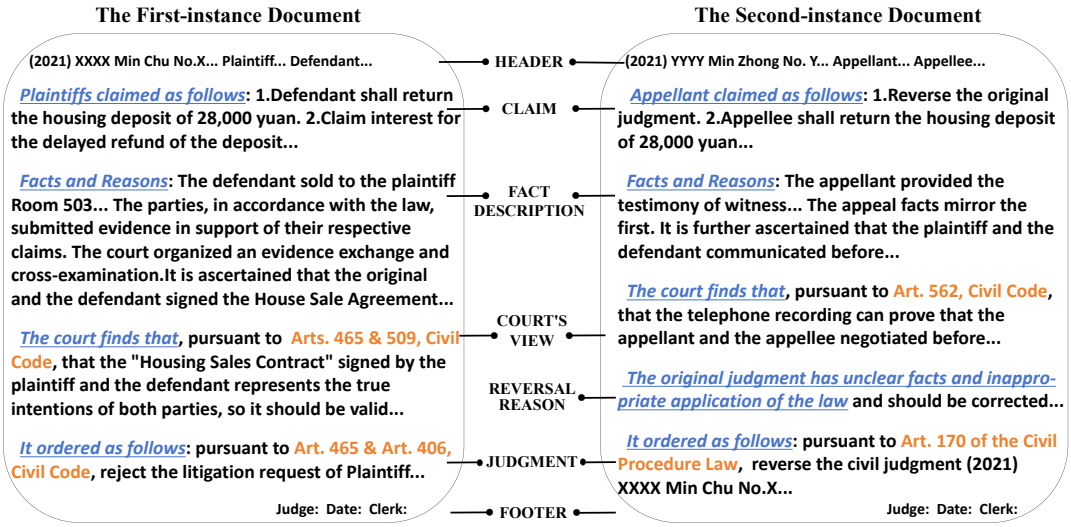

Figure 2: Structural comparison of first-instance and second-instance documents. Underlined phrases represent typical legal expressions used to segment each paragraph in real-world documents.

tive, these demands translate into heightened challenges for LegalAI models: appeal cases involve longer and more interdependent documents, require cross-referencing between trial stages, and call for deeper, multi-step legal reasoning. As a result, appellate tasks are substantially more complex than typical first-instance LegalAI tasks, yet remain underexplored—**particularly due to the lack of standardized datasets, task definitions, and strong evaluation baselines.**

To bridge this gap, we introduce AppealCase, a dataset specifically constructed to support the modeling of civil appellate reasoning. AppealCase contains 10,000 matched pairs of first-instance and second-instance judgment documents collected from China Judgments Online [1], spanning 91 civil causes of action. As a representative of the civil law tradition, China's legal system offers valuable insights for other jurisdictions with similar legal foundations. Notably, 50% of the cases in the dataset result in judgment reversals, providing a balanced perspective on appellate outcomes. To facilitate downstream modeling, we design a dedicated annotation scheme tailored to the appellate setting, capturing structured legal elements.

Building upon AppealCase, we introduce a suite of benchmark tasks that capture the unique challenges of appellate case reasoning and decision-making:

1. **Judgment Reversal Prediction**, which aims to anticipate whether the appellate court will overturn the first-instance decision. This task not only supports appellate adjudication, but also helps identify potential errors in first-instance judgments, offering early diagnostic value.
2. **Provision Relevance Prediction** and **Legal Judgment Prediction**, to assist appellate courts in reviewing and adjudicating appeal cases;
3. **Court View Generation**, to support the drafting of comprehensive appellate judgments.

We conduct a comprehensive evaluation of the appellate scenario using AppealCase, benchmarking 20 non-reasoning, reasoning, and legal domain-specific LLMs on the newly proposed tasks. **Results show that current LLMs struggle to handle the complexities of appellate reasoning**: for instance, all models achieve an average F1 score below 50% on the Judgment Reversal Prediction task. This performance gap underscores the unique challenges posed by appeal cases and highlights the research value of developing dedicated datasets and evaluation protocols for this overlooked yet essential part of the legal system.

The contributions of this paper can be summarized as follows:

1. We establish a comprehensive LegalAI scenario centered on appellate cases, highlighting the legal and technical distinctions between appeals and first-instance trials.
2. We present *AppealCase*, a large-scale dataset of 10,000 matched first-instance and second-instance civil judgment pairs across 91 causes of action. The dataset will be publicly available at https://github.com/ythuang02/AppealCase.
3. We define a suite of benchmark tasks for appellate reasoning and conduct extensive evaluations on 20 LLMs, revealing substantial limitations of current models in this setting.

2 The Dataset for Appellate Case Analysis

We constructed the AppealCase dataset specifically to facilitate research in appellate scenarios, capturing the essential features and structural complexities of appellate judgment documents. Further background and jurisprudential analysis is provided in Appendix A.

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

2.1 Dataset Overview

AppealCase contains 10,000 appellate cases covering 91 civil causes of action. Each case comprises matched pairs of first-instance and second-instance judgments, as illustrated in Figure 2.

Judgment documents in both first-instance and second-instance cases follow a relatively fixed format, typically including the following sections as illustrated in Figure 2, typically includes the following sections: the **header**, which contains metadata such as the case number and court name; the **claim**, which summarizes the plaintiff's or appellant's demands; the **fact description**, presenting the court's account of the case facts, including statements from both parties and relevant evidence; the **court's view**, which explains the legal reasoning and application of law. In appeal cases where the original judgment is modified, an additional section detailing the reasons for reversal is included to justify the appellate court's disagreement with the lower court. The document concludes with the **judgment**, stating the final decision, and the **footer**, which provides the names of the judges and judgment date.

2.2 Annotation Scheme

To support modeling of appellate procedures, each case in AppealCase is also accompanied by five structured annotations designed to capture key aspects of appellate adjudication:

Judgment Reversal A binary label indicating whether the second-instance court overturned the first-instance decision. This annotation supports tasks such as judgment reversal prediction, helping identify potential judicial errors and assess case outcomes under appellate review.

Reasons for Reversal This annotation applies only to cases in which the second-instance court reverses the first-instance decision. According to the Civil Procedure Law of the People's Republic of China, this includes *errors in factual determination* and *errors in the application of law*, or both. Factual errors refer to issues in the trial court's assessment of evidence or understanding of case facts, while legal errors relate to the misapplication or misinterpretation of laws. Further details are provided in Appendix A.2.

Claims A list of individual claims raised in the first-instance proceedings. Each entry records the support status of a specific claim in both the first-instance and second-instance judgments, labeled as *fully supported*, *partially supported*, or *not supported*. This allows for fine-grained analysis of how judicial opinions change across trial levels.

Legal Provisions A list of legal provisions explicitly cited in the second-instance judgment, including the name of the statute and the article numbers. This supports analysis of how appellate courts apply and interpret relevant legal norms.

New Information A binary label indicating whether new evidence were introduced during the appeal. The presence of new information often alters the appellate court's fact-finding and reasoning process, and thus plays an essential role in modeling procedural differences between trial and appeal.

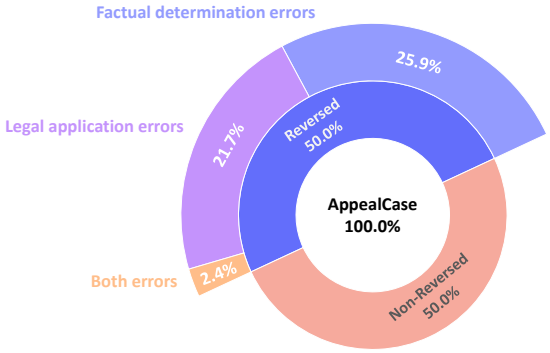

Figure 3: Distribution of reversal reasons in the AppealCase dataset.

Cause of Action	Proportion
Private Lending	13.08%
Labor Dispute	9.94%
Sales Contract	9.86%
Motor Vehicle Traffic Accident	6.36%
Contract	5.28%
Housing Lease Contract	4.20%
Construction Contract	3.50%
Labor Contract	3.20%
Housing Sale Contract	2.74%
Lease Contract	2.18%

Table 1: Proportion of the top-10 causes of action.

Type	Dataset
# Cases	10,000
# Types of Cause of Actions	91
Avg. Number of Claims	2.61
Avg. Number of Legal Provisions	3.13
Avg. Length in Judgment Document	4,243.46
in first-instance	3,672.48
in second-instance	4,818.44

Table 2: Dataset Statistics.

2.3 Descriptive Statistics

Figure 3 shows the distribution of judgment outcomes, with an equal number of reversal and non-reversal cases (50% each). This 1:1 sampling ratio is intentionally designed to facilitate balanced training for reversal prediction tasks. In 2.4% of the cases, the reasons for reversal involve both factual determination errors and legal application errors. Table 1 further breaks down the dataset by cause of action. The most common type of dispute is "private lending," which constitutes 13.08% of all cases. The top 10 causes together account for 60% of the dataset, providing a diverse yet representative set of civil litigation scenarios.

Table 2 summarizes the length characteristics of the judgment documents. Second-instance judgments are, on average, significantly longer than their first-instance counterparts—4,818 vs. 3,672 Chinese characters—reflecting the

Task	Type	# Sample	Metric
Judgment Reversal Prediction	Multi-label Classification	10,000	Macro-averaged Precision, Recall, F1
from the first-instance perspective	Multi-label Classification	5,481	Macro-averaged Precision, Recall, F1
from the second-instance perspective	Multi-label Classification	4,519	Macro-averaged Precision, Recall, F1
Provision Relevance Prediction	Multi-choice Selection	10,000	Subset Accuracy, Samples-averaged Precision, Recall, F1
Legal Judgment Prediction	Single-label Classification	26,143	Accuracy, Macro-averaged Precision, Recall, F1
Court View Generation	Text Generation	10,000	ROUGE-{1, 2, L}, BLEU-{1, 2, 3}, LLM-as-Judger

Table 3: Overview of the AppealCase benchmark tasks, including task types, number of samples, and evaluation metrics.

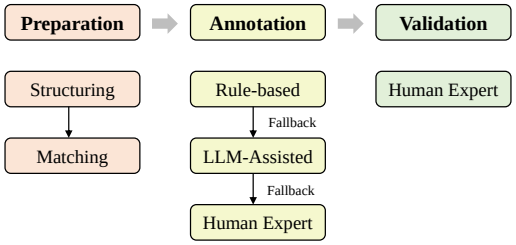

Figure 4: Overview of the three-stage construction pipeline.

added complexity of appellate proceedings. Such complexity arises from the need to reassess factual findings, address legal arguments, and incorporate new evidence.

2.4 Dataset Construction

We construct the AppealCase dataset through a three-stage pipeline as shown in Figure 4: sample preparation, multi-layer annotation, and expert validation. We begin by structuring judgment documents and matching first-instance with second-instance cases. Annotation is conducted in three sequential stages—rule-based, LLM-assisted, and expert annotation. Finally, expert evaluation is performed to assess the overall quality and consistency of the dataset.

Document Structuring We first structured the judgment documents by identifying six core sections—header, claim, fact description, court's view, judgment, and footer—as illustrated in Figure 2. Benefiting from the highly standardized structure and the fixed expressions used in judicial documents, section segmentation can be effectively achieved using keyword-based rules, as detailed in Appendix B.1. Documents that could not be reliably segmented were excluded from the dataset.

Document Matching To establish the mapping between first-instance and second-instance cases, we leveraged the case numbers embedded in the documents. Specifically, we extracted the unique first-instance case number from the judgment section of the second-instance document, typically expressed in phrases like "upholding the judgment of case number X." We then retrieved the corresponding first-instance document to form a matched pair.

Rule-based Annotation We first applied rule-based methods to automatically annotate five types of labels. For example, Judgment Reversal is determined by detecting phrases such as "appeal dismissed" or "original judgment upheld" in the judgment section. The rules used are detailed in Appendix B.2.

LLM-assisted Annotation For cases not covered by rule-based heuristics, we employed a powerful LLM (Qwen-max) to complete the annotation. This applies to the three partially rule-covered schemas: Reasons for Reversal, Claims, and New Information. For each instance, the LLM generated ten responses using sampling with different temperature settings (from 0.1 to 1.0). If all corresponding annotation results were consistent, the result was accepted as the final annotation. The prompt designs are detailed in Appendix B.3.

Human Expert Annotation If the annotation results from the LLM are not completely consistent, the annotation task was escalated to legal experts. These experts—comprising experienced three judges and practicing attorneys—were provided with the same prompts used in the LLM annotation stage. They independently reviewed and discussed the annotation results to reach a consensus, which served as the final golden label. It played a critical role in ensuring the overall quality and reliability of the dataset.

Human Expert Evaluation To further validate the dataset quality, we asked the same group of experts to jointly review 500 randomly selected case pairs. They discussed and annotated each case to produce a unified gold label. The results showed that only 4 cases (0.8%) produced by our multi-layer annotation framework were partial inconsistencies with the gold labels, demonstrating the high accuracy and reliability of the AppealCase dataset.

3 Task Definition

To evaluate the capabilities of models in appellate scenarios, we propose five new benchmark tasks grounded in the AppealCase dataset, as summarized in Table 3. These tasks are designed to capture the unique cognitive and reasoning challenges faced by appellate courts, including: Judgment Reversal Prediction from the perspective of the first-instance and the second-instance, Provision Relevance Prediction, Legal Judgment Prediction, and Court View Generation in the second-instance scenario. We provide task examples and elaborate on the rationale behind the choice of evaluation metrics in Appendix C.

3.1 Judgment Reversal Prediction

The judgment reversal prediction task helps reduce the first-instance misjudgment rate and assists in adjudication in the second-instance. To reflect the differences in information acquisition during actual trials, we divide the data based on whether new information is introduced at the second-instance stage: cases with new information are regarded as

Category	Model	First-instance Perspective			Second-instance Perspective		
		Precision	Recall	F1	Precision	Recall	F1
Non-Reasoning	DeepSeek-V3	44.94	41.87	42.53	55.56	54.97	54.49
	Qwen2.5-72B	47.62	41.23	40.49	55.84	59.45	**57.40**
	LLaMA3.3-70B	40.75	45.22	34.85	50.42	56.94	48.08
	GPT-4.1	**51.93**	36.53	32.58	**60.28**	47.30	44.80
	GLM-4-Air	38.64	42.09	33.24	42.08	41.10	38.72
	Doubao-1-5-pro	42.57	47.42	**44.57**	55.28	59.27	54.73
	Baichuan2-7B	26.00	33.37	26.60	34.99	34.87	25.90
	Qwen2.5-7B	38.28	34.31	30.42	46.02	41.56	40.18
	Llama3.1-8B	34.38	34.34	18.43	41.45	36.06	28.05
Reasoning	DeepSeek-R1	44.17	43.54	43.06	54.03	55.56	54.77
	R1-Distill-Qwen-32B	40.01	45.61	40.58	49.28	57.36	52.07
	QwQ-32B	42.31	**51.41**	40.30	52.38	54.79	49.95
	Qwen3-32B	40.06	49.19	39.30	49.91	59.19	50.64
	GLM-Z1-Air	39.34	43.34	39.46	49.98	54.34	48.90
	GPT-o4-mini	43.67	40.49	40.36	54.31	49.16	47.85
	Grok-3-mini	37.23	49.37	37.41	48.97	**62.64**	53.69
	R1-Distill-Qwen-7B	35.82	44.84	37.71	37.87	52.98	41.68
	Qwen3-8B	39.79	51.34	36.03	49.21	55.44	47.35
Domain	DISC-LawLLM	32.51	33.85	30.05	35.11	34.11	23.09
	Wisdom Interrogatory	32.92	34.20	30.47	33.74	33.93	22.52

Table 4: Performance on judgment reversal prediction from the first-instance and second-instance perspective. Overall best and best results for each category are in **bold** and underline, respectively.

Perspective	Model	Precision	Recall	F1
First-instance	BERT	**58.24**	49.56	52.70
	Qwen3	56.27	**56.93**	**56.10**
Second-instance	BERT	60.03	55.74	56.20
	Qwen3	**67.97**	**63.00**	**63.46**

Table 5: Performance of fine-tuned models on judgment reversal prediction.

the second-instance perspective, while those without are regarded as the first-instance perspective, reflecting the differences in information available to the courts at the two trial levels.

Problem 1 (Judgment Reversal Prediction from the first-instance perspective)**.** *Given the first-instance document and the second-instance claim, the task is to predict the reasons for reversal.*

Problem 2 (Judgment Reversal Prediction from the second-instance perspective)**.** *Given the first-instance document and the second-instance claim and fact description, which contains new information introduced in the second instance, the task is to predict the reasons for reversal.*

3.2 Provision Relevance Prediction

Legal provisions form the foundation of judicial decisions. In the legal provision prediction task, our goal is to accurately select the most relevant legal provisions based on the facts of the case. To achieve this, we prepare 10 candidate options for each case, with the correct option coming from the legal provisions annotated for the case, while the remaining distractors are randomly selected from legal provisions involved in other cases.

Problem 3 (Provision Relevance Prediction)**.** *Given the candidate legal provisions, and the second-instance header, claim, and fact description, the task is to select the relevant legal provisions.*

3.3 Legal Judgment Prediction

Unlike previous judgment prediction tasks (Cui, Shen, and Wen 2023), the task here focuses on adjudication results of each claim from the first-instance at the second-instance stage. We examine whether each claim continues to be supported in the second instance, with results categorized as fully, partially, or not supported. This provides a new perspective for in-depth analyzing changes in claim support across different trial levels.

Problem 4 (Legal Judgment Prediction)**.** *Given the first-instance fact description, the second-instance claim and fact description, and the claim to be judged, the task is to predict whether this claim is supported in the second instance.*

3.4 Court View Generation

Unlike the first-instance court view generation task (Wu et al. 2020), second-instance court view generation not only requires the independent application of law, but also a review of the fact-finding and legal application in the first-instance judgment, and, when necessary, clarification of the specific reasons for reversal. The second-instance court view generation task requires the LegalAI model to reflect the unique perspective and logic of second-instance review and re-judgment, which is of great significance for understanding the supervisory and remedial mechanisms of the judicial trial system.

Problem 5 (Court View Generation)**.** *Given the second-instance claim and fact description, the task is to generate the second-instance court's view, reasons for reversal, and judgment. The summary of the first-instance document is usually already included in the second-instance fact description.*

4 Experiments

We conducted a comprehensive evaluation of the five new LegalAI tasks across 20 models. The experimental details can be found in Appendix D.

Category	Model	Provision Recommendation				Judgment Prediction			
		SA	Precision	Recall	F1	ACC	Precision	Recall	F1
Non-Reasoning	DeepSeek-V3	38.65	87.22	78.00	77.88	64.26	69.08	63.00	61.91
	Qwen2.5-72B	28.27	73.84	82.31	72.24	66.67	67.19	66.39	66.33
	LLaMA3.3-70B	21.58	59.50	46.15	45.21	58.59	59.38	58.67	58.24
	GPT-4.1	27.69	73.18	69.97	65.59	68.19	68.35	68.49	68.19
	GLM-4-Air	25.67	67.96	58.02	55.61	46.50	52.07	47.97	44.19
	Doubao-1-5-pro	**49.55**	**90.50**	**83.00**	**83.65**	**70.93**	**71.49**	**70.34**	**70.62**
	Baichuan2-7B	27.27	48.22	24.77	26.90	43.41	42.64	41.38	38.25
	Qwen2.5-7B	31.02	84.32	67.35	69.44	53.34	59.08	52.48	48.79
	Llama3.1-8B	24.90	49.21	34.80	35.31	42.84	48.38	39.56	36.23
Reasoning	DeepSeek-R1	33.46	84.85	60.66	66.15	63.33	65.95	63.65	62.41
	R1-Distill-Qwen-32B	31.11	76.09	63.08	63.41	62.35	63.34	63.49	62.31
	QwQ-32B	24.27	64.97	83.93	64.05	56.57	59.69	58.35	56.16
	Qwen3-32B	30.16	73.92	75.68	68.46	62.81	62.95	63.31	62.77
	GLM-Z1-Air	30.29	72.62	52.02	56.31	57.16	60.98	58.33	56.75
	GPT-o4-mini	27.98	60.81	47.86	49.30	54.13	60.17	56.51	53.01
	Grok-3-mini	20.48	59.75	63.99	56.02	66.20	66.83	66.07	65.90
	R1-Distill-Qwen-7B	23.65	44.92	31.45	31.16	36.35	42.41	39.17	32.71
	Qwen3-8B	19.35	57.68	62.44	50.06	59.86	60.41	60.30	59.73
Domain	DISC-LawLLM	6.04	22.72	58.56	27.24	34.95	37.45	37.98	29.93
	Wisdom Interrogatory	5.67	23.01	58.91	27.31	34.75	36.15	37.79	29.64

Table 6: Performance on legal provision recommendation and provision relevance prediction.

4.1 Model Categories and Evaluation Setup

We selected a diverse set of pretrained language models as baselines, classified into three categories: non-reasoning models, reasoning models, and domain-specific models, covering representative architectures at different scales. Detailed model providers, model versions, and references can be found in Appendix D.1.

Non-Reasoning Models The open-source models include DeepSeek-V3, Qwen2.5-72B, and LLaMA3.3-70B, while the closed-source models comprise GPT-4.1, GLM-4-Air, and Doubao-1.5-pro. To support low-resource environments, we also introduce three small open-source models: Baichuan2-7B, Qwen2.5-7B, and LLaMA3.1-8B.

Reasoning Models These models possess advanced reasoning capabilities and are designed for complex inference tasks. Open-source models include DeepSeek-R1, R1-Distill-Qwen-32B, QwQ-32B, and Qwen3-32B, while closed-source reasoning models include GLM-Z1-Air, GPT-o4-mini, and Grok-3-mini. Smaller-scale LLMs in this category include R1-Distill-Qwen-7B and Qwen3-8B.

Legal Domain-Specific Models These models are pretrained and fine-tuned on large-scale legal corpora, including DISC-LawLLM (Yue et al. 2023) and Wisdom Interrogatory (ZhihaiLLM 2023).

4.2 Results on Judgment Reversal Prediction

From Table 4, we observe the following results: **1)** All models perform poorly on the judgment reversal prediction task, highlighting its difficulty. Under the first-instance perspective, no model achieves an F1 score above 50%; under the second-instance perspective, while performance improves, the best F1 is only 57.40%, and more than half of the models remain below 50%. **2)** Existing domain-specific models are constrained by limited context windows and struggle to process long, structured judicial documents. **3)** Models perform better in the second-instance perspective, likely due to the inclusion of summarized information from the first-instance trial, which aids reasoning. These results underscore the challenge of factual inconsistency and sparse signal in reversal prediction, especially from the first-instance view.

We also fine-tuned two small-scale models, BERT-base-Chinese (Devlin et al. 2019) and Qwen3-0.6B (Yang et al. 2025), using 80% of the data. As shown in Table 5, while fine-tuning improves performance, the average F1 remains below 60%, suggesting that current model architectures lack the capability to effectively model appellate scenarios.

4.3 Results on Provision Recommendation and Judgment Prediction

From Table 6, we find: **1)** Doubao-1.5-pro achieves strong performance on both provision relevance prediction and legal judgment prediction, with F1 scores of 83.65% and 70.62%, respectively. **2)** However, the Subset Accuracy (SA) for provision prediction is below 50% for all models—including Doubao's 49.55%—indicating that models often miss some relevant provisions despite high overall F1. This reflects incomplete legal reasoning and limited grasp of comprehensive statutory relevance.

4.4 Results on Court View Generation

From Table 7, we observe: **1)** Even the best model, Qwen2.5-72B, achieves average ROUGE and BLEU scores below 40%, revealing a large gap between generated court views and authentic judicial documents in both structure and legal style. **2)** Models pretrained on Chinese corpora show an advantage in generation quality. These results emphasize the difficulty of court view generation, which requires precise legal logic, coherent argumentation, and domain-specific writing style—all of which remain open challenges for current LLMs.

4.5 Case Study on Error Analysis

In the Judgment Reversal Prediction task, there are 401 cases in AppealCase where all 20 models failed to make correct predictions. Notably, all these cases involve judgment reversals, with 83% reversed due to errors in the application of law. We identify the following key challenges in these hard cases:

Category	Model	Court View Generation						
		ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	LLM-Judger
Non-Reasoning	DeepSeek-V3	50.43	24.41	29.50	37.23	26.01	19.65	7.73
	Qwen2.5-72B	**52.66**	**27.61**	**32.21**	40.14	**29.14**	**22.78**	7.67
	LLaMA3.3-70B	41.33	18.30	23.68	23.04	15.33	11.16	7.09
	GPT-4.1	47.40	19.22	24.38	38.65	24.80	17.31	7.85
	GLM-4-Air	41.42	16.85	22.41	23.97	15.27	10.59	7.04
	Doubao-1-5-pro	50.74	25.26	29.81	**40.93**	28.71	21.84	7.78
	Baichuan2-7B	43.70	22.14	26.99	26.71	19.21	15.07	7.18
	Qwen2.5-7B	49.26	24.84	29.65	34.59	24.70	19.08	7.55
	Llama3.1-8B	31.66	12.44	17.11	13.62	8.62	6.09	5.90
Reasoning	DeepSeek-R1	46.98	20.60	26.40	34.62	22.94	16.62	7.79
	R1-Distill-Qwen-32B	46.17	22.38	27.56	27.49	19.39	14.65	7.82
	QwQ-32B	48.32	21.99	27.55	35.75	24.16	17.66	**7.96**
	Qwen3-32B	47.24	20.44	25.70	36.32	23.82	17.11	7.91
	GLM-Z1-Air	44.50	18.38	23.75	33.03	21.15	15.13	7.89
	GPT-o4-mini	42.18	14.94	20.93	29.41	17.52	11.60	7.68
	Grok-3-mini	47.44	20.11	24.92	35.11	22.98	16.33	7.70
	R1-Distill-Qwen-7B	33.78	12.11	18.29	15.24	9.05	5.87	7.11
	Qwen3-8B	46.55	20.05	25.19	34.85	22.69	16.07	7.88
Domain	DISC-LawLLM	29.59	11.25	15.99	11.97	7.38	5.16	6.11
	Wisdom Interrogatory	29.52	11.20	15.91	12.02	7.41	5.18	6.12

Table 7: Performance on court view generation.

- **Fine-grained legal knowledge:** Appeals frequently involve disputes over nuanced legal classifications—such as differentiating between lending and partnership, or employment and contract-for-work relationships. These subtle distinctions are often key to the appellate decision but require detailed legal knowledge that existing models commonly fail to grasp.
- **Dynamic legal knowledge:** Some reversals involve time-sensitive legal standards, such as assessing whether a monthly interest rate exceeds four times the one-year loan market quotation rate. These benchmarks are dynamic, and models struggle to retrieve or interpret the relevant data accurately.
- **Legal reasoning ability:** In appellate proceedings, courts often reassess the division of liability among parties, sometimes adjusting the proportion or legal basis compared to the first-instance decision. This process requires a deeper level of legal reasoning, particularly in interpreting factual findings and justifying modifications. Models generally struggle to trace this type of reasoning, resulting in inaccurate reversal predictions.

These failure cases underscore the fundamental challenges of modeling the appellate scenario. More detailed analysis and case examples are provided in Appendix E.

5 Related Work

5.1 Legal Artificial Intelligence Research

Recent advances in large-scale pre-trained models like BERT and GPT have boosted LegalAI tasks such as provision matching, fact extraction, and judgment prediction (Zhong et al. 2020). Specialized models for long legal texts (Xiao et al. 2021), integrated retrieval-judgment systems (Qin et al. 2024), and knowledge-enhanced prompting for Chinese cases (Sun, Huang, and Wei 2024) have been proposed. Graph-based approaches incorporating domain knowledge, e.g., GraphWordBag for confusing charge prediction (Li et al. 2024a) and the constraint-enhanced GJudge model (Tong et al. 2024), improve prediction accuracy. Causal inference methods enhance consistency and interpretability in European Court of Human Rights cases (Santosh et al. 2022). Despite comprehensive surveys summarizing these advances and challenges (Feng, Li, and Ng 2022), research on complex appellate tasks like second-instance retrieval and court view generation remains limited, as emphasized by CAIL2024 [2].

5.2 Legal AI Benchmarks and Datasets

Current LegalAI benchmarks primarily focus on independent single-stage tasks, with variations in coverage and evaluation depth. Internationally, LegalBench (Guha et al. 2023) based on U.S. federal law and EURLEX (Chalkidis et al. 2020) centered on EU legislation provide extensive classification annotations, serving as foundational resources. Domestically, the annual CAIL evaluation covers a wide range of tasks and serves as the main data source for LawBench (Fei et al. 2024). LexEval (Li et al. 2024d) and LegalAgentBench (Li et al. 2024c) focus on Chinese law but are limited to first-instance judgments, lacking analysis of second-instance reasoning and cross-instance joint evaluation.

6 Conclusion

This paper focuses on the underexplored appellate scenario in LegalAI by constructing the AppealCase dataset, which contains 10,000 realistically paired first-instance and second-instance civil judgments across 91 causes of action. To support modeling the appellate process, we design a scheme specifically tailored to second-instance trials and implement a multi-layer annotation framework that captures key aspects. Based on this dataset, we define five benchmark tasks as baselines for evaluating model capabilities in appellate scenarios. Experimental results on 20 mainstream models show that current LLMs perform poorly on core appellate tasks—especially judgment reversal prediction—highlighting the challenges posed by second-instance reasoning and legal knowledge application. We hope AppealCase will facilitate future research on LegalAI for appel-

[2] http://cail.cipsc.org.cn/task_summit.html?raceID=3&cail_tag=2024

late analysis and contribute to more consistent and accurate judicial decision-making.

Acknowledgment

This work was supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2024C01259, 2025C02037) Key R&D Program of Hangzhou (2025SZDA0254), Ant Group, Chongqing Ant Consumer Finance Co., Ant Group through CCF-Ant Research Fund.

References

Büttner, M.; and Habernal, I. 2024. Answering legal questions from laymen in German civil law system. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2015–2027. St. Julian's, Malta: Association for Computational Linguistics.

Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; and Androutsopoulos, I. 2020. LEGAL-BERT: The Muppets straight out of Law School. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2898–2904. Online: Association for Computational Linguistics.

Cui, J.; Shen, X.; and Wen, S. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11: 102050–102071.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Fei, Z.; Shen, X.; Zhu, D.; Zhou, F.; Han, Z.; Huang, A.; Zhang, S.; Chen, K.; Yin, Z.; Shen, Z.; et al. 2024. LawBench: Benchmarking Legal Knowledge of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7933–7962.

Feng, Y.; Li, C.; and Ng, V. 2022. Legal Judgment Prediction: A Survey of the State of the Art. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 5461–5469.

Guha, N.; Nyarko, J.; Ho, D.; Ré, C.; Chilton, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D.; Zambrano, D.; 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.

Li, A.; Chen, Q.; Wu, Y.; Cai, M.; Zhou, X.; Wu, F.; and Kuang, K. 2024a. From Graph to Word Bag: Introducing Domain Knowledge to Confusing Charge Prediction. arXiv:2403.04369.

Li, A.; Wu, Y.; Liu, Y.; Kuang, K.; Wu, F.; and Cai, M. 2024b. Enhancing Court View Generation with Knowledge Injection and Guidance. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5896–5906. Torino, Italia: ELRA and ICCL.

Li, H.; Chen, J.; Yang, J.; Ai, Q.; Jia, W.; Liu, Y.; Lin, K.; Wu, Y.; Yuan, G.; Hu, Y.; et al. 2024c. LegalAgentBench: Evaluating LLM Agents in Legal Domain. *arXiv preprint arXiv:2412.17259*.

Li, H.; Chen, Y.; Ai, Q.; Wu, Y.; Zhang, R.; and Liu, Y. 2024d. LexEval: A Comprehensive Chinese Legal Benchmark for Evaluating Large Language Models. In *Proceedings of the Thirty-eighth Conference on Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track*.

Merryman, J.; and Pérez-Perdomo, R. 2018. *The civil law tradition: an introduction to the legal systems of Europe and Latin America*. Stanford University Press.

Qin, W.; Cao, Z.; Yu, W.; Si, Z.; Chen, S.; and Xu, J. 2024. Explicitly Integrating Judgment Prediction with Legal Document Retrieval: A Law-Guided Generative Approach. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2210–2220.

Santosh, T.; Xu, S.; Ichim, O.; and Grabmair, M. 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*, 1120–1138.

Sergot, M. J.; Sadri, F.; Kowalski, R. A.; Kriwaczek, F.; Hammond, P.; and Cory, H. T. 1986. The British Nationality Act as a logic program. *Communications of the ACM*, 29(5): 370–386.

Sun, J.; Huang, S.; and Wei, C. 2024. Chinese legal judgment prediction via knowledgeable prompt learning. *Expert Systems with Applications*, 238: 122177.

Tong, S.; Yuan, J.; Zhang, P.; and Li, L. 2024. Legal Judgment Prediction via graph boosting with constraints. *Information Processing & Management*, 61(3): 103663.

Wiratunga, N.; Abeyratne, R.; Jayawardena, L.; Martin, K.; Massie, S.; Nkisi-Orji, I.; Weerasinghe, R.; Liret, A.; and Fleisch, B. 2024. CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering. In *International Conference on Case-Based Reasoning*, 445–460. Springer.

Wu, Y.; Kuang, K.; Zhang, Y.; Liu, X.; Sun, C.; Xiao, J.; Zhuang, Y.; Si, L.; and Wu, F. 2020. De-Biased Court's View Generation with Causality. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 763–780. Online: Association for Computational Linguistics.

Xiao, C.; Hu, X.; Liu, Z.; Tu, C.; and Sun, M. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2: 79–84.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yue, S.; Chen, W.; Wang, S.; Li, B.; Shen, C.; Liu, S.; Zhou, Y.; Xiao, Y.; Yun, S.; Huang, X.; et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

ZhihaiLLM. 2023. Wisdom Interrogatory Model Card.

Zhong, H.; Xiao, C.; Tu, C.; Zhang, T.; Liu, Z.; and Sun, M. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5218–5230.

Zhou, Z.; Shi, J.-X.; Song, P.-X.; Yang, X.-W.; Jin, Y.-X.; Guo, L.-Z.; and Li, Y.-F. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *arXiv preprint arXiv:2406.04614*.