AppellateGen: A Benchmark for Appellate Legal Judgment Generation

Anonymous ACL submission

Abstract

Legal judgment generation is a critical task in legal intelligence. However, existing research in legal judgment generation has predominantly focused on first-instance trials, relying on static fact-to-verdict mappings while neglecting the dialectical nature of appellate (second-instance) review. To address this, we introduce AppellateGen, a benchmark for second-instance legal judgment generation comprising 7,351 case pairs. The task requires models to draft legally binding judgments by reasoning over the initial verdict and evidentiary updates, thereby modeling the causal dependency between trial stages. We further propose a judicial Standard Operating Procedure (SOP)-based Legal Multi-Agent System (SLMAS) to simulate judicial workflows, which decomposes the generation process into discrete stages of issue identification, retrieval, and drafting. Experimental results indicate that while SLMAS improves logical consistency, the complexity of appellate reasoning remains a substantial challenge for current LLMs. The dataset and code are publicly available at: https://anonymous.4open.science/r/AppellateGen-5763.

1 Introduction

The integration of Large Language Models (LLMs) into the legal domain has advanced legal NLP, moving from simple clause retrieval to complex reasoning tasks such as passing the Bar Exam (Katz et al., 2024), and transitioning from discriminative Legal Judgment Prediction (LJP) (Cui et al., 2023; Feng et al., 2022) to Legal Document Generation (Li et al., 2025c; Yue et al., 2023). Despite these advancements, contemporary methodologies predominantly adopt a First-Instance Perspective (Zhong et al., 2018; Liu et al., 2023; Xu et al., 2020), where generative models synthesize judgments directly from static factual descriptions. However, this linear approach neglects a critical phase: the Appellate (Second-Instance) Review (illustrated in **Appendix A**). Distinct from the fact-to-verdict mapping of a first trial, appellate proceedings are inherently dialectical, necessitating a comparative analysis of the original judgment against appellant grievances to rectify errors in fact-finding or legal application (Wechsler, 1977).

This distinction is particularly pronounced in Civil Law jurisdictions. Here, appellate courts operate under the principle of *continuation of trial*, allowing for the re-examination of facts and the admission of new evidence (Apple and Deyling, 1995). This procedural feature necessitates models capable of processing dynamic factual updates. According to judicial statistics from the Supreme People's Court of China, the reversal rate of second-instance cases consistently hovered around 10% to 11% (exceeding 637,000 cases) from 2021 to 2023 (Supreme People's Court of the People's Republic of China, 2021, 2022, 2023). This substantial volume of modified verdicts underscores the complexity inherent in appellate adjudication (Luneburg and Nordenberg, 1981; Bruhl, 2010), which demands complex reasoning to identify discrepancies and produce a binding judgment that affirms, reverses, or remands the initial ruling.

Despite the importance of appellate review, existing legal benchmarks exhibit limitations in **Task Formulation** and **Contextual Granularity**. First, regarding formulation, existing benchmarks primarily assess discriminative capabilities through multiple-choice questions or classification (Chalkidis et al., 2022a; Fei et al., 2024). While recent initiatives have integrated generative tasks (Li et al., 2025b, 2024), they predominantly focus on drafting documents from static facts, overlooking the dialectical nature of appeals. Second, in terms of granularity, current methodologies typically model legal documents as isolated data points. Retrieval-centric benchmarks (Pipitone and Alami, 2024) and document generation datasets (Li et al., 2025c) generally treat case processing as a single-

step inference task. This approach neglects the procedural interdependency between adjudicatory stages, specifically the dependency between initial judgments and appellate reviews.

To bridge this gap, we introduce `AppellateGen`, the first benchmark dedicated to **Second-Instance Legal Judgment Generation**. Unlike previous benchmarks, `AppellateGen` models the complete dispute lifecycle. We curate **7,351 paired cases**, linking first-instance judgments with their corresponding appellate outcomes. We enhance this with a fine-grained annotation schema that identifies points of contention, extracts new facts introduced during the appeal, and categorizes the rationale for reversal.

To validate the challenge posed by `AppellateGen` and investigate robust solutions, we propose the **judicial Standard Operating Procedure (SOP)-based Legal Multi-Agent System (SLMAS)** as a strong baseline. We observe that standard LLMs often struggle when processing the conflicting narratives and long-context evidence inherent in appellate review, leading to hallucinations. To address this, SLMAS imposes a Standard Operating Procedure (SOP) inspired by judicial workflows, decomposing the generation task into discrete stages: *Analysis, Search, Predict, and Write*. By orchestrating agents specialized in issue identification, retrieval, prediction, and drafting, our framework explicitly models the intermediate reasoning states required for valid adjudication. Empirical results demonstrate that this modular approach significantly enhances logical consistency and reversal prediction accuracy, surpassing both commercial and domain-specific baselines.

In summary, our contributions are threefold:

- **Pioneering Task Definition:** We extend the frontier of legal NLP from static first-instance judgment prediction to Second-Instance Legal Judgment Generation. This task necessitates a paradigm shift towards *dynamic dialectical reasoning*, requiring generative models to identify judicial errors and synthesize conflicting narratives between the original verdict and new evidence.
- **The `AppellateGen` Benchmark:** We construct the first large-scale benchmark of 7,351 paired legal cases that explicitly links trial proceedings with appellate outcomes. Equipped with a fine-grained annotation schema for reversal rationales and evidentiary updates, this dataset serves as a rigorous testbed for evaluating long-context reasoning and causal consistency in legal adjudication.
- **Methodological Framework & Insights:** We propose a judicial Standard Operating Procedure (SOP)-based Legal Multi-Agent System (SLMAS). By decomposing the generation process into dispute identification, retrieval, prediction, and drafting, SLMAS has more outstanding generation performance and mitigates logical hallucinations. Furthermore, our extensive evaluation reveals that general-purpose LLMs with strong reasoning capabilities outperform traditional legal domain-specific models on this complex task.

2 Related Work

2.1 Legal Judgment Prediction and Case Document Generation

Legal Judgment Prediction (LJP) has progressed from utilizing hierarchical legal structures (Luo et al., 2017) and few-shot learning (Hu et al., 2018) to deploying attention-based architectures (Chen et al., 2019; Li et al., 2019; Kang et al., 2019) and LLMs (Nigam et al., 2024) for outcome deduction. Despite these advancements, the field has predominantly prioritized prediction over full document generation. While recent initiatives like CaseGen (Li et al., 2025c) have pivoted toward generation, they remain restricted to first-instance trials, neglecting the appellate stage. Within the Chinese judicial framework, second-instance judgments are pivotal for their supervisory role, correcting errors from lower courts and unifying adjudication standards. To address this gap, we introduce the task of Second-instance Legal Judgment Generation.

2.2 Benchmarks in the Legal Domain

Existing benchmarks predominantly focus on isolated, single-stage tasks, varying significantly in scope. On an international scale, foundational resources like LegalBench (Guha et al., 2023) and LexGLUE (Chalkidis et al., 2022a) provide extensive classification annotations, while LEXTREME (Niklaus et al., 2023), FairLex (Chalkidis et al., 2022b), and ArabLegalEval (Hijazi et al., 2024) extend evaluation to multilingual and fairness contexts. In the domain of Chinese law, the CAIL competition series serves as a backbone for recent benchmarks emphasizing cognitive reasoning and syllogism, such as LawBench (Fei et al.,

First Instance	Second Instance
Full Text: *A case concerning a dispute over the right of subrogation between the plaintiff, PICC, and the defendants, Mr. Wen...*	**Full Text:** *Mr. Wen appealed the first-instance judgment in a case concerning a subrogation claim dispute...*
Facts: Wen (fully liable) collided with Han. Han's vehicle was repaired for $8,586. Insurer paid Han and sought recovery from Wen. Wen claimed "old damages" existed but provided no proof.	**New Facts:** Court retrieved police body-cam footage. Footage confirms Han admitted to "old injuries" on site. Damage height didn't match the accident impact.
Legal Articles: - *Insurance Law* Art. 60 - *Road Traffic Safety Law* Art. 76	**Legal Articles:** - *Civil Procedure Law* Art. 170(1)(2)
Judgment: 1. PICC (Insurer 2) to pay $2,000. 2. **Wen (Defendant) to pay $6,586.**	**Judgment:** 1. Affirm PICC's $2,000 payment. 2. **Revoke Wen's $6,586 payment order.**
Is_reversal: `True` | **Reason for Reversal:** Based on newly obtained police footage, the repair scope clearly exceeded the damage caused by the current accident. The trial court's **erroneous fact-finding** regarding the damage amount is corrected, and the excess claim is dismissed.	

Figure 1: An example from our dataset (translated from Chinese). **Facts** refers to the key factual findings extracted from the unstructured full text via a Large Language Model (LLM). **Reason for Reversal** denotes the explanatory rationale synthesized by the LLM, which identifies the critical discrepancies between the first and second instances that led to the judgment change.

2024), LexEval (Li et al., 2024), and LAiW (Dai et al., 2025). Beyond general understanding, recent works have targeted specific system components, including retrieval (LegalBench-RAG (Pipitone and Alami, 2024), LegalSearchLM (Kim et al., 2025)), instruction tuning (LawInstruct (Niklaus et al., 2025)), and agentic workflows (LegalAgentBench (Li et al., 2025b)). While generative capabilities are partially addressed by CaseGen (Li et al., 2025c) and GreekBarBench (Chlapanis et al., 2025), these tasks often lack the dialectical depth of appellate review. A notable exception is the recent AppealCase dataset (Huang et al., 2025), which explicitly targets second-instance scenarios. However, this work focuses on the discriminative classification task, failing to capture the complex reasoning required to draft appellate opinions.

2.3 Legal Multi-Agent Systems

Multi-Agent Systems (MAS) have addressed single-agent limitations such as hallucination through collaborative simulation (Ji et al., 2023; Li et al., 2023). In the legal domain, frameworks like ChatLaw (Cui et al., 2023) and logic-graph-enhanced models (Yuan et al., 2026) leverage such collaboration to optimize first-instance judgment prediction and consultation. While these systems efficiently decompose tasks via role-playing, they are inherently designed for mapping facts to verdicts rather than the dialectical reasoning required for appellate review. Second-instance generation demands a comparative analysis to weigh original judgments against appeal grounds.

3 `AppellateGen` Construction

3.1 Data Collection and Legal Instances Matching

We collected 4,706,987 raw legal judgment documents spanning from 2021 to 2024[1] from *China Judgments Online*[2], the official government repository for Chinese judicial documents. Since the raw repository stores judgments as independent documents without explicit links between first-instance and second-instance proceedings, we developed a multi-stage pipeline to match the same cases in different instances (Details illustrated in **Appendix B.1**). To assess the quality of the matching process, we randomly sampled 5% of the paired cases for manual expert review, achieving a consistency rate of approximately 96%. After the whole process, all

[1]This specific timeframe was selected to ensure legal consistency, as the Civil Code and the 11th Amendment to the Criminal Law of the People's Republic of China were enacted in 2021, providing a stable statutory basis for the judgments in our dataset.

[2]`https://wenshu.court.gov.cn/`

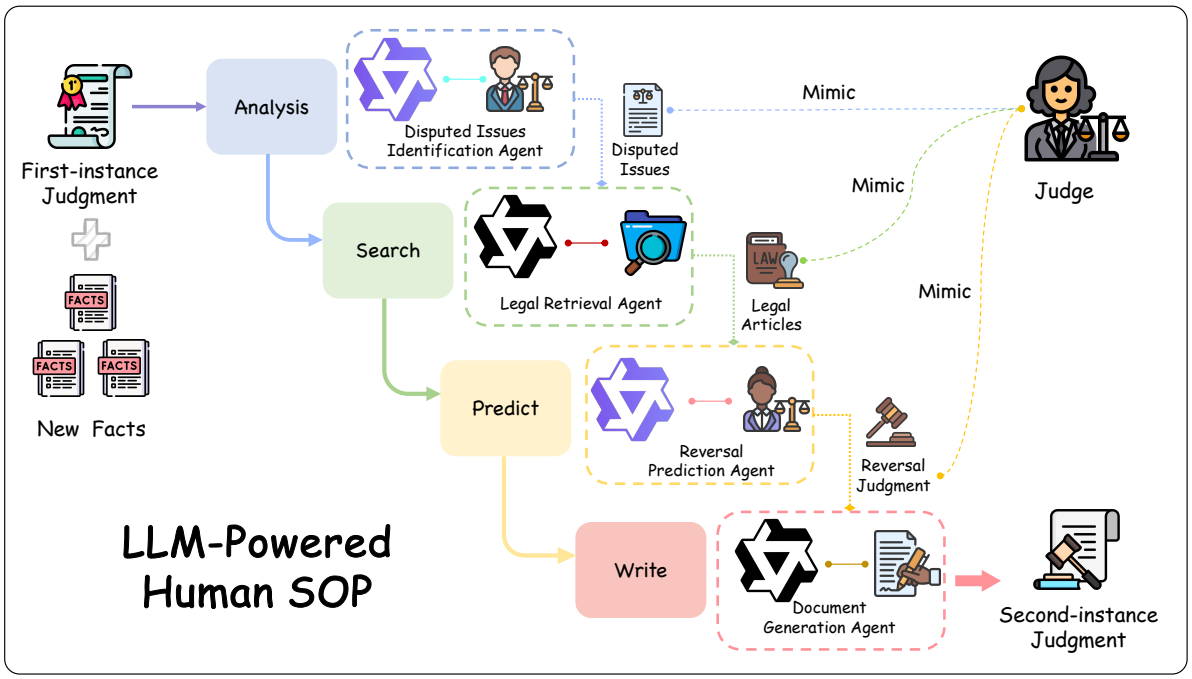

Figure 2: SOP-based Legal Multi-Agent System (SLMAS). With first-instance judgment and new facts provided, the framework mimics the human judicial Standard Operating Procedure (SOP) by orchestrating four specialized agents. These agents sequentially identify disputed issues, retrieve relevant legal statutes, predict the reversal outcome, and finally draft the appellate judgment, ensuring the generation is grounded in explicit legal reasoning.

the documents have been de-identified to protect the privacy of litigants. We show an example of our dataset in **Figure 1**.

3.2 Dataset Statistics

We initially collected approximately 4 million raw legal documents through the pipeline described above. A final dataset comprising 7,351 high-quality case pairs is curated, this sparsity is attributed to three main factors: 1) Natural Filtering: A significant portion of cases conclude at the first instance without appeal; 2) Procedural Outcomes: In the appellate stage, a high volume of cases are resolved through mediation or withdrawal, which typically yield brief procedural rulings rather than the substantive judgments required for reasoning tasks; 3) Strict Quality Control: To ensure the reliability of the benchmark, we enforced rigorous matching criteria based on case numbers and litigant names, discarding pairs with incomplete chains of custody or insufficient textual content. We report the key descriptive statistics in **Table 1**.

3.3 Fine-grained Data Annotation

While standard metrics (e.g., ROUGE (Lin, 2004)) assess lexical overlap, they fail to capture the legal validity and logical consistency required for the task of second-instance document generation. To address this limitation, we designed a fine-grained annotation schema for the matched case pairs. These annotations establish a rigorous semantic ground truth that enables multi-dimensional evaluation beyond surface-level text generation.

Metric	1st Instance	2nd Instance
Count	7,351	7,351
Avg. Length	5,004.1	5,752.8
Max. Length	40,219	41,732
Min. Length	916	879
Label Dist.	*Rev.: 10.7%*	*Aff.: 89.3%*

Table 1: Descriptive statistics of the dataset. Lengths are measured in Chinese characters. The dataset contains 7,351 pairs in total (10.7% Reversal, 89.3% Affirmation).

3.3.1 Annotation Schema

Unlike previous datasets that focus solely on static judgment outcomes, our schema captures the *dynamic interaction* and *adjudicative evolution* between the two instances. This allows us to assess whether the generated documents accurately reflect the legal reasoning shifts from the first to the second trial. The annotations were obtained via

a hybrid framework combining LLM-driven automated extraction with strict human verification. The schema comprises four key dimensions:

Judgment Result Classification: This binary label serves as the high-level outcome indicator. We classify the second-instance ruling as either *Affirm* (maintaining the original verdict) or *Reverse* (encompassing cases remanded for retrial or where the judgment was directly altered). This provides the fundamental stance for the generated text.

Rationale for Reversal: For cases involving a reversal, we identify the specific legal trigger. This label distinguishes whether the appellate decision was driven by *erroneous fact-finding*, *misapplication of law*, or *procedural violations* in the first instance. This annotation is critical for evaluating if the generated document correctly explains the *why* behind the verdict.

New Evidence and Facts Identification: This label summarizes newly ascertained facts derived exclusively from the second-instance judgment, explicitly excluding facts established in the first instance. This design requires the LLM to synthesize the first-instance factual summary with these novel appellate details to perform legal reasoning, thereby simulating authentic judicial practice.

Statutory Citation Extraction: This label targets the specific legal articles cited by the appellate court as the controlling authority for its decision. We extract the full citation (e.g., *Article X of the Civil Code*), distinguishing operative laws that dictate the ruling from those merely referenced in the procedural history.

3.3.2 Automated Annotation Pipeline

To guarantee the reliability of annotations within complex legal contexts, we established a robust three-stage pipeline:

Structural Decomposition: We employed customized regular expressions to partition unstructured judgment documents into distinct functional units, specifically the *Factual Description* and *Judicial Reasoning* sections. This segmentation mitigates input noise and directs the model's attention toward contextually relevant information.

LLM-Driven Annotation: We used DeepSeek-V3.2 (Liu et al., 2025) to generate preliminary annotations. To bolster reasoning transparency and label accuracy, we incorporated a Chain-of-Thought (CoT) prompting strategy (Detailed in Appendix B.2). This approach compels the model to articulate the underlying legal rationale and calculate a self-assessed confidence score (γ) prior to determining the final label.

Human-in-the-Loop Verification: To ensure the construction of a rigorous ground truth, we instituted a confidence-aware review protocol. Instances where the model's confidence fell below a designated threshold ($\gamma < 0.8$) were flagged for manual adjudication by legal experts.

4 SOP-based Legal Multi-Agent System

4.1 Problem Formulation

We formalize the task of Second-Instance Legal Judgment Generation as a conditional text generation problem involving multi-hop reasoning. Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ denote the dataset. For each case, the input context $\mathbf{x}$ comprises a tuple $\mathbf{x} = \langle \mathbf{d}_{\text{1st}}, \mathbf{e}_{\text{new}} \rangle$, where $\mathbf{d}_{\text{1st}}$ represents the textual content of the first-instance judgment (including initial fact-finding and verdict), and $\mathbf{e}_{\text{new}}$ denotes the new evidentiary facts introduced during the appeal (where $\mathbf{e}_{\text{new}} = \emptyset$ if no new evidence is presented).

The target output is the second-instance judgment document, denoted as $\mathbf{y} = (y_1, y_2, \ldots, y_T)$. Our objective is to learn a parameterized model θ that maximizes the log-likelihood of generating $\mathbf{y}$ conditioned on $\mathbf{x}$. Unlike standard end-to-end generation, we posit that the generation of $\mathbf{y}$ strictly depends on a sequence of intermediate reasoning states $\mathbf{z} = \{\mathbf{z}_{\text{issue}}, \mathbf{z}_{\text{law}}, \mathbf{z}_{\text{verdict}}\}$. Thus, the objective function is formulated as:

$$\mathcal{L}(\theta) = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \log P_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) \qquad (1)$$

where $\mathbf{z}$ is the latent rationales derived from the judicial Standard Operating Procedure (SOP).

4.2 Framework Overview

Directly mapping $\mathbf{x} \rightarrow \mathbf{y}$ often leads to hallucinations and logical inconsistencies due to the complexity of appellate review. To address this, we propose a judicial Standard Operating Procedure (SOP)-based Legal Multi-Agent System (SLMAS) (Shown in **Figure 2**) that mimics the human judicial Standard Operating Procedure (SOP) (DeMarco and Lister, 2013; Hong et al., 2023).

We model the appellate process as a deterministic assembly line executed by four specialized agents:

$$\mathcal{A} = \{\mathcal{A}_{\text{issue}}, \mathcal{A}_{\text{retr}}, \mathcal{A}_{\text{pred}}, \mathcal{A}_{\text{write}}\}. \quad (2)$$

Each agent $\mathcal{A}_k$ acts as a functional operator that transforms the current state and prior contexts into a structured intermediate output. The overall workflow forms a directed acyclic graph (DAG) of reasoning (Digitale et al., 2022), ensuring that the final generation is grounded in explicit legal logic rather than statistical correlations.

4.3 Sequential Reasoning Workflow

The execution of our framework follows a sequential chain corresponding to the four stages of human adjudication: *Analysis, Search, Predict, and Write*.

Stage 1: Disputed Issues Identification (Analysis). The process begins with the *Disputed Issues Identification Agent* ($\mathcal{A}_{\text{issue}}$). This agent analyzes the discrepancy between the first-instance judgment and the new evidence to distill the core points of contention. Formally, given the input $\mathbf{x}$, the agent extracts a set of disputed issues $\mathbf{z}_{\text{issue}}$:

$$\mathbf{z}_{\text{issue}} = \mathcal{A}_{\text{issue}}(\mathbf{d}_{\text{1st}}, \mathbf{e}_{\text{new}}) \quad (3)$$

where $\mathbf{z}_{\text{issue}}$ serves as a structured summary of the appeal grounds, guiding the subsequent retrieval process.

Stage 2: Legal Retrieval (Search). To ground the reasoning in statutory law, the *Legal Retrieval Agent* ($\mathcal{A}_{\text{retr}}$) utilizes the identified issues $\mathbf{z}_{\text{issue}}$ to query an external legal knowledge base $\mathcal{K}$ (Details Illustrated in Appendix C). Unlike generic retrieval, this agent focuses on statutes specifically relevant to the controversy:

$$\mathbf{z}_{\text{law}} = \mathcal{A}_{\text{retr}}(\mathbf{z}_{\text{issue}}, \mathcal{K}) \quad (4)$$

Here, $\mathbf{z}_{\text{law}}$ represents the retrieved relevant legal articles and judicial interpretations, providing the normative basis for the review.

Stage 3: Reversal Prediction (Predict). Before drafting the document, the *Reversal Prediction Agent* ($\mathcal{A}_{\text{pred}}$) mimics a judge's internal deliberation. This agent integrates the case context and derived legal grounds to adjudicate the appeal, specifically predicting whether the primary verdict warrants affirmation or reversal. The agent generates a predictive rationale $\mathbf{z}_{\text{verdict}}$:

$$\mathbf{z}_{\text{verdict}} = \mathcal{A}_{\text{pred}}(\mathbf{d}_{\text{1st}}, \mathbf{e}_{\text{new}}, \mathbf{z}_{\text{issue}}, \mathbf{z}_{\text{law}}) \quad (5)$$

This step acts as a logical checkpoint, ensuring the final document's conclusion is consistent with the analyzed facts and laws.

Stage 4: Judgment Generation (Write). Finally, the *Document Writing Agent* ($\mathcal{A}_{\text{write}}$) generates the final second-instance judgment. Conditioned on the entire reasoning chain $\mathbf{z} = \{\mathbf{z}_{\text{issue}}, \mathbf{z}_{\text{law}}, \mathbf{z}_{\text{verdict}}\}$, the agent produces the document token-by-token:

$$P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) = \prod_{t=1}^{T} P_\theta\big(y_t \mid y_{<t}, \mathbf{d}_{\text{1st}}, \mathbf{e}_{\text{new}}, \mathbf{z}_{\text{issue}}, \mathbf{z}_{\text{law}}, \mathbf{z}_{\text{verdict}}\big) \quad (6)$$

By explicitly conditioning on $\mathbf{z}$, the framework ensures that the generated text strictly adheres to the identified issues, retrieved laws, and predicted verdict, effectively mitigating hallucination.

5 Experiment

5.1 Experiment Settings

Test Dataset Construction. We curated a dedicated test set through stratified sampling. We then obtained 1,000 distinct cases from the original dataset, consisting of 500 cases where the initial judgment was reversed and 500 cases where it was affirmed. This 1:1 ratio ensures that the evaluation metrics accurately reflect the model's reasoning ability rather than the label distribution.

Baseline. We leverage LLMs ranging from tens of billions to hundreds of billions of parameters as baselines. The models include: (1) General Open-source Models: gpt-oss-120b (Agarwal et al., 2025) and Qwen3 (Yang et al., 2025); (2) Commercial Close-source Models: Gemini 2.5 (Comanici et al., 2025); (3) Legal-specific Models: Domain-adapted models including DISC-LawLLM (Yue et al., 2023) and Wisdom Interrogatory (Wu et al., 2024).

Implementation Details. All agents in SLMAS are instantiated using Qwen3 (Yang et al., 2025), with role-specific prompts and tool to enable functional specialization. For reproducibility, we fix the decoding temperature at 0.1 across all trials. To ensure fair comparison, identical prompts are employed, and input sequences are truncated to fit within each model's respective context window.

5.2 Automatic Evaluation

We employ a hybrid framework combining LLM-as-a-Judge (reference-free expert scoring) (Li et al.,

Model	Standard Metrics		LLM-as-a-Judge Scores					Reversal Prediction Accuracy
	ROUGE-L	BERTScore	Verdict Consistency	Fact Consistency	Legal Application	Logical Reasoning	Average	
Single Non- Reasoning Agent								
Qwen3-235B-A22B-Instruct-2507	0.344	0.866	2.136	2.098	2.503	2.291	2.257	61.42%
Qwen3-30B-A3B-Instruct-2507	0.360	0.863	2.144	2.322	2.504	2.120	2.273	60.43%
Qwen3-8B	0.313	0.840	1.564	2.077	2.151	1.469	1.815	57.92%
Gemini-2.5-Flash-Lite	0.521	0.836	1.955	2.314	2.411	1.713	2.098	57.58%
Gemini-2.5-Flash	**0.536**	0.856	2.396	2.451	2.558	2.323	2.432	66.53%
Single Reasoning Agent								
QwQ-32B	0.277	0.851	2.366	**2.541**	2.689	2.243	2.460	60.06%
Qwen3-30B-A3B-Thinking-2507	0.288	0.836	2.213	2.373	2.606	2.000	2.298	64.81%
DeepSeek-R1-0528-Qwen3-8B	0.316	0.837	2.137	2.229	2.293	1.697	2.089	57.87%
gpt-oss-120b	0.320	0.797	1.262	2.078	2.059	1.749	1.787	41.81%
Single Domain Agent								
DISC-LawLLM	0.080	0.660	1.643	2.075	2.237	1.683	1.910	35.22%
Wisdom Interrogatory	0.333	0.721	1.443	1.983	1.731	1.240	1.599	37.34%
Multi Agents (Our SLMAS)								
Qwen3-235B-A22B-Instruct-2507	0.337	**0.878**	2.353	2.472	**2.705**	**2.615**	**2.536**	**67.24%**
Qwen3-30B-A3B-Instruct-2507	0.353	0.834	**2.460**	2.406	2.430	2.274	2.393	66.49%
Qwen3-8B	0.272	0.810	1.712	2.181	2.308	1.684	1.971	61.27%

Table 2: Performance comparison on Appellate Legal Judgment Generation task. **Bold** indicates best performance.

2025a) and standard reference-based metrics to assess the legal validity of generated judgments.

LLM-as-a-Judge. We utilize DeepSeek-V3.2 (Liu et al., 2025) to score generations on a Likert scale (0–5) based on expert-designed rubrics (Joshi et al., 2015). The detailed rubrics are located in **Appendix D**. The evaluation encompasses four dimensions: (1) Verdict Consistency: Checks if the ruling direction (Affirm/Reverse) matches the ground truth and whether the judgment results are the same. A mismatch results in a nullified score ($S_{verdict} = 0$). (2) Fact Consistency: Assesses the reconstruction of facts and handling of evidence. (3) Legal Application: Verifies the accurate citation of operative statutes. (4) Logical Reasoning: Evaluates syllogistic rigor regarding appeal grounds.

Standard Metrics. Complementing the judge model, we report ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2019) to measure lexical overlap and semantic similarity against reference judgments.

5.3 Main Result

5.3.1 Effectiveness of the SLMAS Framework

Table 2 demonstrates that the SLMAS framework consistently enhances performance across models of varying scales. Using Qwen3-235B-A22B-Instruct-2507 as a primary example, SLMAS improves the average generation quality score from 2.257 to 2.536 and increases Reversal Prediction Accuracy from 61.42% to 67.24%. This performance exceeds that of the proprietary Gemini 2.5 Flash (66.53%), indicating that open-source models can achieve state-of-the-art results when equipped with effective agentic strategies. The improvement is attributed to the structural decomposition of the legal syllogism. By decoupling issue identification from the drafting process, SLMAS effectively models the judicial workflow, thereby mitigating the hallucinations and logical inconsistencies observed in long-context generation tasks.

5.3.2 Reasoning Capabilities versus Domain Knowledge

A significant divergence in performance is observed between general-purpose LLMs and domain-specific models. While general models achieve accuracies exceeding 60%, traditional legal models such as DISC-LawLLM and Wisdom Interrogatory lag significantly, with accuracies between 35% and 37%. This disparity suggests that appellate judgment generation relies less on static knowledge retrieval and more on comparative reasoning capabilities. Domain-specific models, trained on static legal corpora, lack the cognitive flexibility required to process dynamic evidentiary conflicts between first and second-instance trials. In contrast, general LLMs, particularly those enhanced with chain-of-thought processes, demonstrate superior adaptability in these logic-intensive scenarios.

5.3.3 Persistent Challenges in Generation Quality

Despite the improvements introduced by SLMAS, generating second-instance judgments remains a complex challenge. Although prediction accuracy

Agent Configuration				Standard Metrics		LLM-as-a-Judge Scores					Reversal Prediction Accuracy
Disp. Issues	Legal Retr.	Rev. Pred.	Doc. Gen.	ROUGE-L	BERTScore	Verd. Cons.	Fact Cons.	Legal App.	Logical Reas.	Average	
✓	✓	✓	✓	0.272	0.810	**1.712**	**2.181**	**2.308**	**1.684**	**1.971**	**61.27%**
✓	✗	✓	✓	0.278	0.820	1.694	2.064	2.085	1.644	1.872	59.30%
✗	✓	✓	✓	0.269	0.803	1.663	2.129	2.191	1.651	1.909	60.41%
✗	✗	✓	✓	**0.331**	0.823	1.621	2.118	2.131	1.614	1.871	58.05%
✗	✗	✗	✓	0.313	**0.840**	1.564	2.077	2.151	1.469	1.815	57.92%

Table 3: We perform the Ablation study using Qwen3-8B. We report Generation Similarity (ROUGE-L) and (BERTScore), LLM-as-a-Judge Scores (including Verdict Consistency, Fact Consistency, Legal Application and Logical Reasoning), and Reversal Prediction Accuracy. The Average column represents the mean of the LLM-as-a-Judge Scores. Best results are **bold**.

has reached approximately 67%, the generation quality scores reveal a persistent gap, with average scores hovering around 2.5 out of 5. The primary bottleneck is identified in the *Logical Reasoning* metric (2.615), where models occasionally fail to construct closed-loop arguments that rigorously connect new evidence to statutory conclusions. This indicates that while instruction tuning improves adherence to format, future advancements may require paradigms specifically designed to enhance long-chain causal reasoning.

5.4 Ablation Study

To evaluate the contribution of individual components, we conducted an ablation study using Qwen3-8B on the test dataset. The results, presented in **Table 3**, highlight the interdependency of reasoning modules and the divergence between surface-level metrics and semantic validity.

5.4.1 Dependency of Retrieval on Issue Identification

The full system achieves optimal performance with an average score of 1.971 and a reversal prediction accuracy of 61.27%. A critical finding is the impact of the *Disputed Issues Identification* agent. Removing this module degrades performance not only in reasoning but also in legal application. Without explicit issue definition, the subsequent retrieval module lacks directional focus, leading to the retrieval of statutes that fail to ground the final judgment. This confirms that SLMAS prevents error propagation by ensuring that each stage is conditioned on a verified intermediate state.

5.4.2 Divergence between N-gram and Semantic Metrics

An analysis of the *No Agents* baseline reveals a discrepancy between traditional metrics and judge-based evaluations. This configuration yields the highest ROUGE-L score (0.331) but a mediocre LLM-as-a-Judge score (1.871). This anomaly exposes the limitation of surface-level metrics in legal generation tasks. The baseline model tends toward verbatim repetition of the first-instance judgment, resulting in high lexical overlap but low dialectical value. Conversely, SLMAS promotes logical restructuring to address appeal grounds, which naturally reduces n-gram overlap with the source text while significantly improving *Verdict Consistency* (1.712 vs. 1.621) and *Logical Reasoning*. This underscores the necessity of employing semantic-aware evaluation over traditional overlap metrics for high-level reasoning tasks.

6 Conclusion

We present `AppellateGen`, a benchmark shifting the focus of legal intelligence from judgment prediction to the dialectical complexity of second-instance judgment generation. Comprising 7,351 annotated case pairs, it evaluates model reasoning over conflicting narratives and judicial errors. Our proposed SLMAS framework, which structurally decomposes the judicial workflow, significantly enhances logical consistency and reversal accuracy. Notably, general LLMs equipped with reasoning capabilities outperform domain-specific models, underscoring the primacy of cognitive flexibility over static domain pre-training for this task. While a gap remains between generated outputs and human expertise, `AppellateGen` establishes a foundation for future research into causal dependency modeling and robust comparative reasoning architectures.

Limitations and Future Work

Our study presents a novel benchmark for second-instance legal judgment generation. However, we acknowledge two primary limitations. First, the geographic and linguistic scope of `AppellateGen` is confined to Mainland China and the Chinese language. This jurisdiction-specific focus, while necessary for legal precision, limits the direct applicability of our models to other legal systems or multilingual contexts. Second, a significant performance bottleneck persists. Although our proposed SLMAS outperforms most of the baselines, the absolute performance remains suboptimal, with a SOTA metric of 67.24% and a LLM-as-a-judge evaluation score of only 2.536 out of 5. These results suggest that current paradigms struggle to fully capture the intricate dialectical reasoning required for high-stakes judicial writing. To address these challenges, we look forward to the emergence of foundation models possessing superior legal reasoning capabilities.

Ethics Statement

The judgment documents of our dataset are all collected from a publicly accessible platform, China Judgments Online, which is widely used in legal AI research (Huang et al., 2025; Li et al., 2025c). We strictly adhere to the data usage policies and redistribution terms of the source platform, ensuring that our release is confined to de-identified textual content within the boundaries of public access. However, legal data contains sensitive information. To strictly protect the privacy of the individuals involved, we have applied rigorous anonymization preprocessing. All Personally Identifiable Information (PII), such as litigants' names, ID numbers, home addresses, and phone numbers, has been desensitized prior to model training and dataset release. We also acknowledge the potential risks associated with algorithmic bias. As the dataset reflects historical judicial decisions, it may inherently contain biases related to gender, region, or specific crime types. Models trained on this data might perpetuate these biases. Therefore, we emphasize that this dataset and the proposed method are intended for academic research and as assistive tools for legal professionals only. They should never be used as a substitute for human judges or to automate final judicial decisions. We strongly advise future researchers to conduct fairness evaluations when deploying systems based on this dataset.

GenAI Usage Statement

We clarify the use of generative AI in this study as follows: (1) Research Methodology: As detailed in Sections 3.3 and 5.2, LLMs were utilized for data annotation and as a scoring judge to assess judgment generation. (2) Writing and Coding: LLMs were used as productivity tools to optimize the codebase and refine the linguistic expression of the paper. All final outputs were critically reviewed by the authors.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

James G Apple and Robert P Deyling. 1995. *A primer on the civil-law system*. Federal Judicial Center.

Aaron-Andrew P Bruhl. 2010. Deciding when to decide: How appellate procedure distributes the costs of legal change. *Cornell L. Rev.*, 96:203.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022b. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. *arXiv preprint arXiv:1908.11521*.

Odysseas S Chlapanis, Dimitrios Galanis, Nikolaos Aletras, and Ion Androutsopoulos. 2025. Greekbarbench: A challenging benchmark for free-text legal reasoning and citations. *arXiv preprint arXiv:2505.17267*.

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

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

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. LAiW: A Chinese legal large language models benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10738–10766, Abu Dhabi, UAE. Association for Computational Linguistics.

Tom DeMarco and Tim Lister. 2013. *Peopleware: productive projects and teams*. Addison-Wesley.

Jean C Digitale, Jeffrey N Martin, and Medellena Maria Glymour. 2022. Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*, 142:264–267.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction: A survey of the state of the art. In *IJCAI*, pages 5461–5469.

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

Faris Hijazi, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem AlZahrani, Hebah AlShamlan, Omar Knio, and George Turkiyyah. 2024. ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 225–249, Bangkok, Thailand. Association for Computational Linguistics.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th international conference on computational linguistics*, pages 487–498.

Yuting Huang, Meitong Guo, Yiquan Wu, Ang Li, Xiaozhong Liu, Keting Yin, Changlong Sun, Fei Wu, and Kun Kuang. 2025. Appealcase: A dataset and benchmark for civil case appeal scenarios. *arXiv preprint arXiv:2505.16514*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

Liangyi Kang, Jie Liu, Lingqiao Liu, Qinfeng Shi, and Dan Ye. 2019. Creating auxiliary representations from charge definitions for criminal charge prediction. *arXiv preprint arXiv:1911.05202*.

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

Chaeeun Kim, Jinu Lee, and Wonseok Hwang. 2025. Legalsearchlm: Rethinking legal case retrieval as legal elements generation. *arXiv preprint arXiv:2505.23832*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2025b. LegalAgentBench: Evaluating LLM agents in legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2322–2344, Vienna, Austria. Association for Computational Linguistics.

Haitao Li, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:25061–25094.

Haitao Li, Jiaying Ye, Yiran Hu, Jia Chen, Qingyao Ai, Yueyue Wu, Junjie Chen, Yifan Chen, Cheng Luo, Quan Zhou, and 1 others. 2025c. Casegen:

A benchmark for multi-stage legal case documents generation. *arXiv preprint arXiv:2502.17943*.

Shang Li, Hongli Zhang, Lin Ye, Xiaoding Guo, and Binxing Fang. 2019. Mann: A multichannel attentive neural network for legal judgment prediction. *IEEE Access*, 7:151144–151155.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. Mlljp: Multi-law aware legal judgment prediction. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 1023–1034.

William V Luneburg and Mark A Nordenberg. 1981. Specially qualified juries and expert nonjury tribunals: Alternatives for coping with the complexities of modern civil litigation. *Va. L. Rev.*, 67:887.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. *arXiv preprint arXiv:1707.09168*.

Shubham Kumar Nigam, Aniket Deroy, Subhankar Maity, and Arnab Bhattacharya. 2024. Rethinking legal judgement prediction in a realistic scenario in the era of large language models. *arXiv preprint arXiv:2410.10542*.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

Joel Niklaus, Lucia Zheng, Arya D. McCarthy, Christopher Hahn, Brian M Rosen, Peter Henderson, Daniel E. Ho, Garrett Honke, Percy Liang, and Christopher D Manning. 2025. LawInstruct: A resource for studying language model adaptation to the legal domain. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 127–152, Albuquerque, New Mexico. Association for Computational Linguistics.

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

RanKKI and LawRefBook Contributors. 2025. Lawrefbook/laws. https://github.com/LawRefBook/Laws.

Supreme People's Court of the People's Republic of China. 2021. Judicial Statistics Bulletin of National Courts in 2021. http://gongbao.court.gov.cn/Details/a6c42e26948d3545aea5419fa2beaa.html. (in Chinese).

Supreme People's Court of the People's Republic of China. 2022. Judicial Statistics Bulletin of National Courts in 2022. http://gongbao.court.gov.cn/Details/20587eaef248beb61ed6596018865c.html. (in Chinese).

Supreme People's Court of the People's Republic of China. 2023. Judicial Statistics Bulletin of National Courts in 2023. http://gongbao.court.gov.cn/Details/a3e86176b272dc94a05d9cb012c2d5.html. (in Chinese).

Herbert Wechsler. 1977. The appellate jurisdiction of the supreme court: Reflections on the law and the logistics of direct review. *Wash. & Lee L. Rev.*, 34:1043.

Yiquan Wu, Yuhang Liu, Yifei Liu, Ang Li, Siying Zhou, and Kun Kuang. 2024. wisdominterrogatory. Available at GitHub.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. *arXiv preprint arXiv:2004.02557*.

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

Weikang Yuan, Kaisong Song, Zhuoren Jiang, Junjie Cao, Yujie Zhang, Chengyuan Liu, Jun Lin, Ji Zhang, Kun Kuang, and Xiaozhong Liu. 2026. A multi-agent framework with legal event logic graph for multi-defendant legal judgment prediction. *Information Processing & Management*, 63(1):104319.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and 1 others. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3540–3549.

A Background on Appellate (Second-Instance) Review in Civil Law Jurisdictions

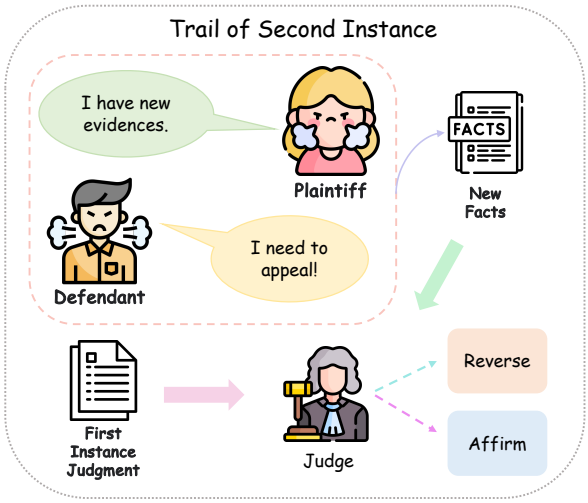

Figure 3: Proceedings of the Second Instance: Following the filing of an appeal by the appellant, the admission of new evidence may result in either the reversal of the verdict or the affirmation of the original judgment.

The procedural scope of appellate review exhibits fundamental differences between Common Law and Civil Law jurisdictions. In Common Law systems, appellate courts typically exercise deference to the lower court regarding factual determinations, restricting their review primarily to errors of law. Conversely, Civil Law jurisdictions adopt a comprehensive approach where the court of second instance functions as a continuation of the trial process, retaining the authority to re-examine both legal applications and factual findings. This study focuses specifically on the Civil Law framework, utilizing a dataset derived from Chinese legal proceedings.

Within this context, the second-instance trial serves as a critical remedial mechanism. As illustrated in **Figure 3**, the procedure is initiated when a litigant, who may be either the plaintiff or the defendant, files an appeal expressing dissatisfaction with the first-instance judgment. A defining characteristic of this process in China is the admissibility of new evidence. Unlike systems that strictly limit the appellate record, this framework permits the introduction of supplementary facts that were previously unavailable or overlooked.

Consequently, the input for the appellate judge comprises two distinct streams of information: the text of the original first-instance judgment and the set of newly presented evidentiary facts. The presiding judge is tasked with synthesizing these inputs to evaluate the validity of the appeal. This review process necessitates a comparative analysis where the judge assesses whether the new evidence sufficiently contradicts the original findings or if the initial application of the law remains robust. The proceeding culminates in a binary judicial determination. The court may decide to affirm the original judgment if the appeal lacks sufficient merit, effectively affirming the prior ruling. Alternatively, if the combination of the original context and new facts reveals substantive errors, the court issues a reversal, thereby altering the legal verdict. **Figure 3** provides a schematic representation of this workflow, delineating the causal path from the filing of the appeal and the admission of evidence to the final adjudication.

Regular Expression Pattern for Instances Matching

```
[ (〔\d{4}[〕) )]\s*[\u4e00-\u9fa5\d]{1,15}\s*(?:民|刑|行|商|知|破|
    执|赔|法|[\u4e00-\u9fa5])?\s*[初终]字?第?\s*\d+\s*号
```

Figure 4: RE pattern for legal instances matching. This pattern can extract the case identifier of first instance from second-instance judgment document.

B AppellateGen Construction Details

To construct AppellateGen, we perform two main steps: (1) Data Collection and Legal Instance Matching, and (2) Fine-grained Data Annotation. The construction details are as follows.

B.1 Legal Instances Matching

First, we employed high-precision Regular Expressions (RE) to extract unique case identifier of first instance from second-instance judgment document. The specific extraction pattern is detailed in **Figure 4**. Subsequently, we retrieved the corresponding first-instance judgment documents based on these identifiers to construct candidate case pairs.

To ensure the semantic integrity of these pairs, filtering out potential mismatches caused by clerical errors in case numbers, we utilized a LLM as a semantic validator. Specifically, we prompted the LLM to compare the litigants, the cause of

action, and the court hierarchy described in both documents, outputting a binary verification score indicating whether they refer to the identical legal dispute. The used prompts are detailed below.

You are an experienced expert in reviewing Chinese legal documents, specializing in the filing and correlation of civil, criminal, and administrative case files. You possess exceptional logical analysis and attention to detail.
Your task is to determine whether two given legal documents (a first-instance judgment and a second-instance judgment) belong to the same case's proceedings (i.e., whether the second instance appealed against the first-instance judgment).
Please follow these steps for logical reasoning, without directly providing a conclusion:
1. Extract Key Information:
From the first-instance judgment: Extract the first-instance case number, plaintiff/defendant's name (or title), and cause of action.
From the second-instance judgment: Extract the second-instance case number, appellant/respondent's name, and the original trial case number cited in the text.
2. Compare Relevance:

- Case Number Verification: Check the "Case Origin" or "Original Trial Circumstances" section...
- Verification of Parties: Does the appellant/appellee in the second instance correspond to the plaintiff/defendant...
- Verification of Facts: Are the core facts of the dispute consistent?

3. Elimination of Interference: Pay attention to distinguishing between "related cases"...
Output Format: Please strictly follow the following JSON format for outputting results:
{
"is_same_case": true/false,
"confidence_score": "0-10",
"reasoning": ""
}

B.2 Fine-grained Data Annotation

Here is the prompt template for LLM-driven annotation.

Please act as a professional expert in Chinese legal document analysis. I will provide the full text of the [First Instance Judgment] and the [Second Instance Judgment] for the same case.
Your tasks are:

1. **Deep Analysis and Extraction:** Compare the two documents and extract key information.
2. **Formatted Output:** Integrate all analysis results into a strict JSON object.

Please output strictly according to the following JSON structure and do not include Markdown tags:

```
{
   "reasoning_trace":         "Your
step-by-step analysis here...",
   "confidence_score": float,
   "is_reversal": bool,
   "first_instance": {
     "facts": "string",
     "disputed_issues": "string",
     "legal_articles": ["string"],
     "judgment": "string"
   },
   "second_instance": {
     "new_facts": "string",
     "disputed_issues": "string",
     "legal_articles": ["string"],
     "judgment": "string"
   },
   "reason_for_reversal": "string"
}
```

Here are the explanations of the following JSON schemas.

- **is_reversal**: A boolean value indicating whether the second-instance judgment reverses.
- **first_instance**: A structured summary of the first-instance judgment, including factual findings, disputed issues, cited legal articles, and the final ruling.

- **facts**: Core facts established in the first instance.
- **disputed_issues**: Key disputes addressed by the court.
- **legal_articles**: Statutory provisions relied upon.
- **judgment**: Outcome of the first-instance ruling.

- **second_instance**: A structured summary of the appellate judgment, emphasizing changes relative to the first instance.
  - **new_facts**: Newly identified facts or evidence.
  - **disputed_issues**: Disputed issues considered on appeal.
  - **legal_articles**: Controlling statutory provisions.
  - **judgment**: Outcome of the second-instance ruling.
- **reason_for_reversal**: Legal grounds for reversal; left empty if the judgment is affirmed.

Please begin analyzing the document content I provide.

C SLMAS details

We established a specialized legal knowledge base derived from statutes promulgated after 2021, utilizing the open-source `LawRefBook` toolkit (RanKKI and LawRefBook Contributors, 2025). A key advantage of this repository is its capability to maintain synchronization with the authoritative *National Laws and Regulations Database of China*[3]. To enable efficient semantic matching, the textual content of every legal article was embedded using Qwen3-Embedding (Zhang et al., 2025).

Upon receiving a search query, the retriever performs a coarse-grained search to identify the top-10 most similar candidates. To ensure precision, an LLM then functions as a fine-grained selector, filtering the retrieved results to pinpoint the specific legal articles and judicial interpretations applicable to the dispute.

[3]This platform serves as the official legislative database maintained by the Ministry of Justice of the People's Republic of China (`https://flk.npc.gov.cn`).

D Evaluation Details

In this section, we provide the specific scoring rubrics and the prompt template used for the LLM-as-a-Judge evaluation described in **Section 5.2**.

Scoring Rubrics

We designed fine-grained criteria to guide the evaluator (DeepSeek-V3.2) in scoring the generated judgments. The scoring scale ranges from 0 to 5 across four specific dimensions. The detailed rubrics are presented in **Table 4**.

Evaluation Prompt Template

We employed a zero-shot prompting strategy with detailed instructions to ensure the LLM evaluator acts as an impartial judge. The full prompt used in our experiments is presented below.

System Role: You are a senior legal expert in the Chinese judicial system, specializing in appellate review. Your task is to evaluate a second-instance judgment.
Evaluation Task: Please score the generated output from 0 to 5 on the following dimensions based on the provided rubrics.
1. Verdict Consistency (0-5): Does the model make the correct decision (Affirm vs. Reverse)? *CRITICAL RULE:* If the ruling direction contradicts the Ground Truth (e.g., model affirms while truth reverses), you **MUST** give 0 points.
2. Fact Consistency (0-5): Does the model accurately describe the facts? Crucially, does it capture the "New Evidence" introduced in the second instance? If the model ignores new evidence that leads to a reversal, score low.
3. Legal Application (0-5): Are the cited statutes correct and relevant? Does it cite the operative substantive law rather than just procedural descriptions?
4. Logical Reasoning (0-5): Is the reasoning coherent? Does it correctly refute or support the appeal grounds? For reversal cases, does it clearly explain *why* the first instance was wrong?
Output Format: Please output a strictly valid JSON object:
{
"verdict_score": int,
"fact_score": int,

```
"law_score": int,
"logic_score": int,
"justification": "A brief explanation of the scoring..."
}
```

Dimension	Score	Criteria Description
1. Verdict Consistency	**0**	**Incorrect Ruling Direction.** The generated judgment contradicts the ground truth (e.g., affirming instead of reversing). **Strict Penalty**
	1–2	**Ambiguous/Incomplete.** The direction is unclear, contradictory, or missing the operative part entirely.
	3–4	**Direction Correct but Flawed.** The general ruling (Affirm/Reverse) is correct, but specific execution orders (e.g., amounts, specific behaviors) are imprecise or missing.
	5	**Perfect Match.** The operative part is identical in legal effect and scope to the ground truth, fully resolving the dispute.
2. Fact Consistency	0–1	**Hallucination.** Fabricates facts not present in the input or contradicts the core timeline of the case.
	2–3	**Missing New Evidence.** Accurately recaps the first-instance facts but fails to identify or incorporate the "New Facts" introduced during the appeal.
	4	**Accurate.** Captures both original facts and new evidence, with minor omissions in non-critical details.
	5	**Comprehensive.** Perfectly reconstructs the factual basis, highlighting the contradiction between new evidence and original findings.
3. Legal Application	0–1	**Erroneous Citation.** Cites non-existent statutes or applies laws completely irrelevant to the case cause.
	2–3	**Incomplete Citation.** Misses the key operative article that dictates the reversal/affirmation, citing only procedural rules.
	4–5	**Precise Citation.** Correctly identifies and applies the specific articles (e.g., Civil Code Art. X) used in the ground truth.
4. Logical Reasoning	0–1	**Incoherent.** The reasoning is logically fractured or fails to address the appellant's grievances entirely.
	2–3	**Mechanical Mapping.** Performs simple fact-to-law mapping but lacks comparative analysis. Misses the causal link between new evidence and the verdict change.
	4	**Strong Syllogism.** Explicitly discusses why the first-instance judgment was erroneous based on the points of contention.
	5	**Expert Dialectics.** Demonstrates a clear "review-and-correct" logic, forming a watertight argument that mirrors a human judge's opinion.

Table 4: Detailed Scoring Rubrics for Automatic Evaluation. The evaluator scores each generated judgment from 0 to 5 based on these criteria. Note that a score of 0 in Verdict Consistency is strictly enforced if the ruling direction is wrong.