

AR-Bench: Benchmarking Legal Reasoning with Judgment Error Detection, Classification and Correction for Chinese Criminal Cases

Yifei Li¹, Richong Zhang^{1*}, Wanyu Tu¹, Zhijie Nie¹,
Haokun Luo¹, Chuantao Yin²

¹CCSE, School of Computer Science and Engineering, Beihang University, Beijing, China

²Sino-French Engineer School, Beihang University, Beijing, China

{liyifei, zhangrc}@act.buaa.edu.cn

Abstract

Legal judgments may inadvertently contain errors, yet human appellate review faces immense pressure to improve efficiency. While current legal AI focuses on legal judgment prediction or document generation, judgment review is fundamentally a retrospective diagnostic task. In this paper, we introduce **APPELLATE REVIEW** to assess large language models' (LLMs) legal diagnostic reasoning underlying appellate review. We specifically target Chinese criminal cases, under the assumption that their highly structured logic, which maps explicit facts to strict statutory penalties, provides a verifiable ground truth. We decompose APPELLATE REVIEW into three subtasks: error detection, error classification, and error correction. To support this, we construct **AR-BENCH**, a controlled benchmark comprising 8,700 judgments with counterfactually introduced errors spanning charges, prison terms, and fines. Evaluations of 16 LLM variants reveal substantial limitations: models classify known errors relatively well but struggle with unguided detection and precise correction. Further analyses of a representative model show that it identifies explicit textual contradictions but fundamentally fails at holistic deductive reasoning and rule-based arithmetic.

1 Introduction

Equality before the law is a fundamental principle of modern justice, crucial for ensuring balanced sentencing in criminal law. Although judicial rulings undergo rigorous procedures, judges may still make errors during the adjudication process due to the complexity of cases and the abstract nature of legal norms (Hu, 2024). To counter this, both civil (Hoquet-McKee, 1981; Woo, 1989; Lindemann and Lienau, 2020) and common (Orfield, 1936; Hessick and Hessick, 2008) law systems have

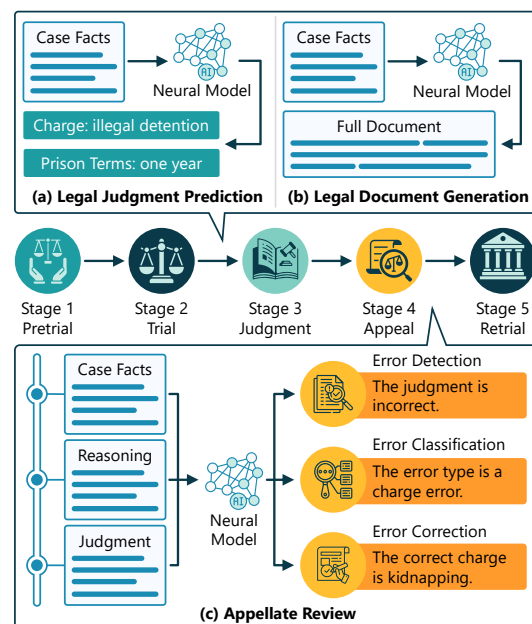


Figure 1: Workflow of legal AI tasks across Chinese judicial stages. While (a) legal judgment prediction and (b) legal document generation assist drafting in pretrial/trial stages, our proposed (c) APPELLATE REVIEW focuses on finalized judgments through three subtasks: error detection, classification, and correction.

established appellate review mechanisms to reexamine the accuracy of legal applications and the appropriateness of sentencing, thereby correcting potential errors in finalized legal judgments.

Taking China as an example, while these appellate review mechanisms are vital for enhancing judicial fairness and curbing discretionary abuses (Bai, 2006), they face immense practical pressures (Wallace, 2005; Zuo, 2018). A 65.8% surge in appeal cases from 2015 to 2024 has significantly burdened judges (PRC, 2025), potentially diminishing the thoroughness of human reviews and error correction. Against this backdrop, a critical question arises: *Can artificial intelligence (AI) be introduced to assist in the review of finalized judgments?*

Figure 1 illustrates that while existing legal AI

*Corresponding author

Benchmark	Legal Facts	Reasoning Process	Law Articles	Sentencing Factors	Attributes Classes	Type of Cases	Jurisdiction
CAIL2018 (Xiao et al., 2018)	✓	✗	✗	✗	2	Criminal	Mainland China
Auto-Judge (Long et al., 2019)	✓	✗	✗	✗	2	Civil	Mainland China
SwissJP (Niklaus et al., 2021)	✓	✗	✗	✗	2	Generic	Switzerland
LBOX OPEN (Hwang et al., 2022)	✓	✗	✗	✗	4	Generic	Korean
AppealCase (Huang et al., 2025)	✓	✓	✗	✗	12	Civil	Mainland China
JuDGE (Su et al., 2025)	✓	✓	✓	✗	8	Criminal	Mainland China
AR-BENCH (Ours)	✓	✓	✓	✓	20	Criminal	Mainland China

Table 1: Comparison between AR-BENCH and existing legal judgment prediction benchmarks in civil law systems.

research has predominantly focused on legal judgment prediction (Xiao et al., 2018; Feng et al., 2022) and document generation (Polsley et al., 2016; Shen et al., 2022; Su et al., 2025), these tasks differ fundamentally from appellate review, particularly within the highly structured Chinese judicial system that this study focuses on. First, they target fundamentally different stages of the legal process. Legal judgment prediction and legal document generation tasks focus on the period before a judgment is rendered, aiming to predict outcomes or draft documents. In contrast, appellate review focuses on the period after a judgment is rendered, examining existing judgments to determine whether to initiate a retrial.

A further distinction lies in their learning objectives. Legal judgment prediction involves predicting legal outcomes from fact descriptions, constituting a classification or prediction task. Appellate review, however, identifies potential errors within finalized judgments, falling under error identification or anomaly detection tasks (Felice and Briscoe, 2015; Yang et al., 2025b). Similarly, while document generation produces structured text from inputs, appellate review deconstructs and diagnoses existing texts, aligning more closely with fact-checking (Thorne et al., 2018; Schlichtkrull et al., 2023) paradigms. While recent benchmarks have extended prediction (Huang et al., 2025) and generation (Yang et al., 2026) tasks to civil appeals, an evaluation benchmark tailored specifically for logical diagnostic review remains underexplored.

To address this research gap, we propose a novel task, **APPELLATE REVIEW**, designed to specifically evaluate models’ ability to identify errors in legal application within legal judgments. We decompose this task into three subtasks: error detection, error classification, and error correction. Given an initial case fact description and a corresponding judgment, models are tasked with (1) detecting whether the judgment contains legal errors;

(2) identifying the specific type of error; and (3) generating a corrected, legally valid revision based on the identified error. To support this, we construct **AR-BENCH**¹, a novel benchmark derived from real-world cases. Unlike the prior datasets shown in Table 1, AR-BENCH provides 20 fine-grained attributes per sample across 8,700 criminal judgments containing manually constructed errors.

A natural question arises as to why AR-BENCH is restricted exclusively to criminal judgments rather than encompassing civil or administrative cases. This restriction is a deliberate choice to ensure objective evaluability. Criminal law operates on a highly structured logical chain: explicit case facts map to specific charges, which dictate strict prison terms and fine intervals. This allows us to define a relatively objective and verifiable ground truth for error correction, enabling precise continuous evaluation metrics. In contrast, civil and administrative appeals frequently involve proportional liability, discretionary judgments, and subjective interpretations (Huang et al., 2025), making their ground truth less amenable to empirical verification than our criminal error constructs. By isolating criminal law, AR-BENCH avoids evaluation noise and provides a rigorous testbed for legal diagnostic reasoning underlying appellate review.

We summarize our contributions as follows:

- **Task Formulation:** We propose APPELLATE REVIEW, a novel task assessing LLMs’ diagnostic reasoning on finalized legal judgments. We structure this task into three closely related subtasks: error detection, error classification, and error correction.
- **Novel Benchmark:** We construct AR-BENCH, a large-scale, fine-grained benchmark derived from Chinese criminal cases with injected anomalous judgments to evaluate models’ diagnostic ability under controlled settings.

¹Data: <https://github.com/Yifei6672/AR-Bench>

- **Empirical Insights:** By evaluating 16 diverse LLM variants on AR-BENCH, we reveal critical limitations in existing models’ ability to handle diagnostic legal logic, underscoring the need for more robust diagnostic reasoning abilities.

2 Related Work

2.1 Legal Judgment Prediction

Legal judgment prediction has undergone a technological evolution from rule-based/statistical methods (Lawlor, 1963; Ulmer, 1963) and traditional machine learning (Liu and Hsieh, 2006; Aletras et al., 2016) to deep learning and pre-trained models (Feng et al., 2022; Wu et al., 2023; Zhang et al., 2023; T.y.s.s et al., 2024; Yue et al., 2024a), with its research objective remaining consistent: predicting legal outcomes based on case fact descriptions. Recent models further enhance their sensitivity to subtle factual distinctions through contrastive learning (Liu et al., 2022; Zhang et al., 2023; Gan et al., 2023) methods and by modeling legal knowledge using graph structures (Bi et al., 2023; Li et al., 2024a; Xu et al., 2024). Although these models have made significant progress in outcome prediction, it remains an open question (An et al., 2022) whether they have genuinely mastered the underlying logic of legal reasoning. Unlike legal judgment prediction, which focuses on predicting judgments based on fact descriptions, our task aims to identify potential legal application errors in final judgments from finalized judgment documents. This requires models to possess deeper legal logical analysis and diagnostic reasoning capabilities.

2.2 Legal Reasoning Benchmarks

As shown in Table 1, conventional legal judgment prediction datasets (Xiao et al., 2018; Hu et al., 2018; Chalkidis et al., 2019) lack fine-grained annotations to systematically evaluate models’ legal reasoning ability. With the rise of LLMs, a series of legal reasoning benchmarks have emerged, covering both comprehensive assessments (Guha et al., 2023; Fei et al., 2024; Li et al., 2024b; Fan et al., 2026) and specialized task (Niklaus et al., 2025; Li et al., 2025) evaluations, encompassing legal judgment prediction (Shui et al., 2023; Zhang et al., 2024), legal document retrieval (Upadhy and T.y.s.s, 2025; Seo et al., 2025), legal document generation (Shen et al., 2022; Su et al., 2025) and legal question answering (Zhong et al., 2020; Liu et al., 2024). However, these benchmarks rarely

offer a systematic evaluation of judicial diagnostic capabilities. To bridge this gap, we introduce APPELLATE REVIEW, aiming to assess a model’s ability to identify potential legal application errors in existing judgments. While concurrent works explore civil appellate proceedings through outcome prediction (Huang et al., 2025) or open-ended document generation (Yang et al., 2026), AR-BENCH uniquely frames APPELLATE REVIEW as a diagnostic task. By strictly focusing on the highly structured logic of criminal law rather than highly contextualized civil disputes, we avoid the evaluation noise associated with generative metrics and establish an objectively verifiable ground truth.

3 APPELLATE REVIEW and AR-BENCH

3.1 Task Formulation

The proposed APPELLATE REVIEW focuses on the legal diagnostic components of appellate review, namely error detection, classification, and correction of legal application errors. The three subtasks are evaluated independently to prevent upstream errors from confounding the downstream evaluation.

Error Detection. Let D denote the judgment document. The model is required to identify the presence of errors in the application of law within D . We frame this as a prompt-constrained binary classification task:

$$\hat{y} = \arg \max_{y \in \{\text{true}, \text{false}\}} P(y \mid D, p_{\text{det}}), \quad (1)$$

where prompt p_{det} restricts the model to output a binary answer $\hat{y} \in \{\text{true}, \text{false}\}$ where true indicates the judgment is correct, and false indicates the judgment contains an error.

Error Classification. Given an erroneous judgment document D , the model is required to classify the specific type of the primary error. We frame this as a prompt-constrained classification task:

$$\hat{E} = \arg \max_{E \in \mathcal{E}} P(E \mid D, p_{\text{cls}}), \quad (2)$$

where prompt p_{cls} strictly restricts the LLM’s output to a target space $\mathcal{E}$ containing three predefined error categories: charge, prison term, and fine, as shown in Figure 2.

Multiple errors often co-occur as cascading anomalies. To streamline statistical evaluation, we label each case by its root cause and restrict the prediction to the primary error type, treating derived errors as consequences of the primary violation.

Error Correction. Let $\mathcal{D}$ denote the judgment document and $E \in \mathcal{E}$ denote the identified error type. The model is required to generate a local correction $\hat{C}$ to replace the original erroneous element. To prevent open-ended evaluation ambiguity, we frame this as a prompt-constrained generation task:

$$\hat{C} = \arg \max_{C \in \mathcal{Y}_E} P(C \mid D, E, p_{\text{cor}}), \quad (3)$$

where prompt p_{cor} strictly restricts the LLM’s output to a structured target space $\mathcal{Y}_E$ containing only the exact charge or penalty value.

3.2 Evaluation Metrics

For error detection and classification, we employ Accuracy (Acc), Macro-Precision (MaP), Macro-Recall (MaR), and Macro-F1 (MaF1) to handle category imbalance. For error correction, the parsed output $\hat{y}$ from $\hat{C}$ is evaluated according to the type of the target variable. For discrete targets (Charge), we use the aforementioned classification metrics against the ground-truth y .

We evaluate continuous numerical targets (Prison Term and Fine) from two perspectives. First, we map the continuous values into predefined legal interval labels to compute the standard classification metrics. Second, because interval-based classification can obscure meaningful differences near boundaries, we additionally utilize distance-aware continuous metrics: ImpScore (Zhong et al., 2018b) and Acc@0.1 (Chen et al., 2019).

ImpScore is defined as the difference between the predicted targets and the ground-truth targets. Assume that the ground-truth target of the i -th case is t_i and the predicted result is $\bar{t}_i$. Then, we define the difference d_i as:

$$d_i = |\log(t_i + 1) - \log(\bar{t}_i + 1)|. \quad (4)$$

After that, the score function $f(v)$ is defined as:

$$f(v) = \begin{cases} 1.0, & \text{if } v \leq 0.2, \\ 0.8, & \text{if } 0.2 < v \leq 0.4, \\ 0.6, & \text{if } 0.4 < v \leq 0.6, \\ 0.4, & \text{if } 0.6 < v \leq 0.8, \\ 0.2, & \text{if } 0.8 < v \leq 1, \\ 0.0, & \text{if } 1 < v. \end{cases} \quad (5)$$

Then the final score should be:

$$S = \sum_{i=1}^M \frac{f(d_i)}{M}. \quad (6)$$

The Acc@0.1 defines a 10% tolerance margin, formulated as:

$$\text{Acc@0.1} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(|\hat{y}_i - y_i| \leq 0.1 \cdot y_i), \quad (7)$$

where $\hat{y}_i$ is the predicted value, y_i is the ground truth, and the indicator function $\mathbb{I}(\cdot)$ assesses whether the prediction falls within this range.

3.3 Data Processing and Base Annotation

The original judgments of AR-BENCH are derived from a publicly available legal document dataset (Su et al., 2025). In designing the benchmark, we explicitly avoid introducing real appellate judgments. Such cases frequently involve factual confounders, such as new evidence, that complicate the evaluation of pure legal reasoning. We opt to inject synthetic errors into first-instance cases to isolate logical errors, ensure a balanced error taxonomy, and prevent LLM data contamination.

We applied rigorous filtering criteria to retain only single-defendant cases. This deliberate choice reduces potential evaluation noise, ensures the atomic nature of synthesized errors, and strictly aligns with our objective of maintaining verifiable evaluation metrics. Regular expressions were employed to extract and standardize the required fields (charges, prison terms, fines, etc.) into a unified format, as they are typically written in a consistent structure in the documents.

To ensure the quality of extracted annotations, we conducted a two-stage manual evaluation. In the first stage, three legal experts manually reviewed a stratified sample of 400 cases. Based on the identified errors, we refined the regular expressions to systematically correct all mislabeled instances. In the second stage, the experts performed a post-validation using a newly and independently sampled set of 400 cases. This final evaluation yielded a Fleiss’ κ of 0.88 and a pass rate of 95.2%.

3.4 Anomalous Judgment Synthesis

Unlike legal judgment prediction tasks, appellate review requires models to diagnose logical fallacies within finalized documents. To this end, we manually injected anomalous logic into standard judgment documents via text modification, synthesizing anomalous judgment samples encompassing six fine-grained error types (E1-E6) as shown in Figure 2. Designed as a diagnostic stress-test rather than reflecting empirical frequencies, these errors

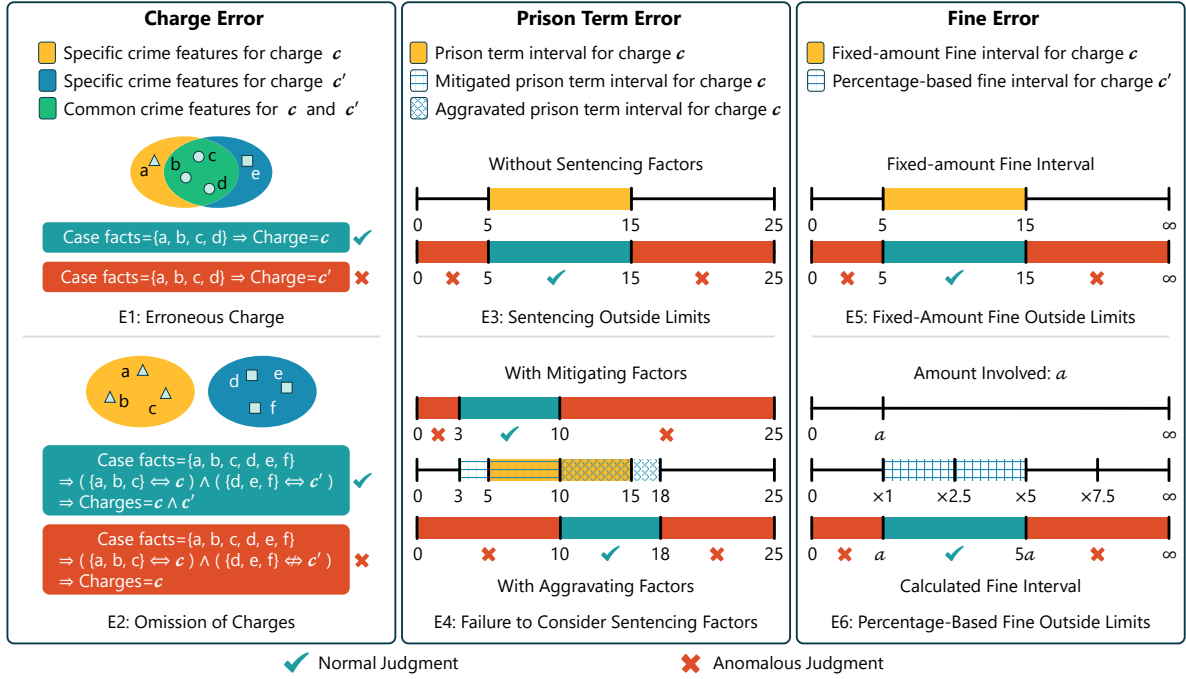


Figure 2: Taxonomy of Six Manually Constructed Error Types for AR-BENCH. This diagram illustrates the designed errors targeting three adjudication elements: **Charges** (E1: Erroneous Charge, E2: Omission of Charges), **Prison Terms** (E3: Sentencing Outside Limits, E4: Failure to Consider Sentencing Factors), and **Fines** (E5: Fixed-amount Fine Outside Limits, E6: Percentage-based Fine Outside Limits).

comprehensively cover three core types: Charges, Prison Terms, and Fines.

Charge Error. Erroneous conviction (E1) refers to situations where two charges are easily confused in a judgment (Bai, 2006) due to highly similar criminal characteristics. Omission of charges (E2) refers to situations where a case should involve multiple offenses, but the initial judgment only recognized the primary offense corresponding to the main criminal characteristics while overlooking other offenses matching different characteristics.

Prison Term Error. Sentencing outside limits (E3) refers to sentences imposed that exceed the statutory sentencing limits specified in criminal law. Failure to consider sentencing factors (E4) refers to a sentencing determination that deviates from the sentencing range established based on such factors, as illustrated in the lower half of Figure 2.

Fine Error. Fine errors encompass both Fixed-Amount Fine Outside Limits (E5) and Percentage-based Fine Outside Limits (E6). Fixed-amount fine outside limits (E5) occurs when the Criminal Law specifies a fixed fine range (from 5 to 15 as shown in Figure 2), and a judge imposes a fine exceeding this range. Percentage-based fine outside limits (E6) occurs when the Criminal Law establishes a

fine range in proportional terms (from $\times 1$ to $\times 5$ as shown in Figure 2), and a judge fails to accurately determine the amount involved during sentencing, resulting in a judgment error where the imposed fine exceeds the proportional range.

Guaranteeing Logical Coherence. To prevent models from cheating by exploiting superficial textual artifacts, we explicitly avoided simple token-level substitutions of specific legal terms, such as final prison terms or charges.

Instead, we implemented a deeper, semantic-level manipulation. For instances such as *Erroneous Conviction* and *Omission of Charges*, we retrieved cases sharing identical key sentencing factors from a single-charge database, and carefully swapped the reasoning sections based on the matched cases. This approach is designed to preserve the local logical coherence of the synthesized documents, forcing models to rely on deep legal reasoning to discover fundamental conflicts between the finalized judgment and the case facts.

Furthermore, based on this rigorous synthesis mechanism, we automatically derived the correct attributes (i.e., the original labels prior to modification) for each anomalous judgment. This automated derivation provides a relatively objective and consistent evaluation baseline for the subsequent

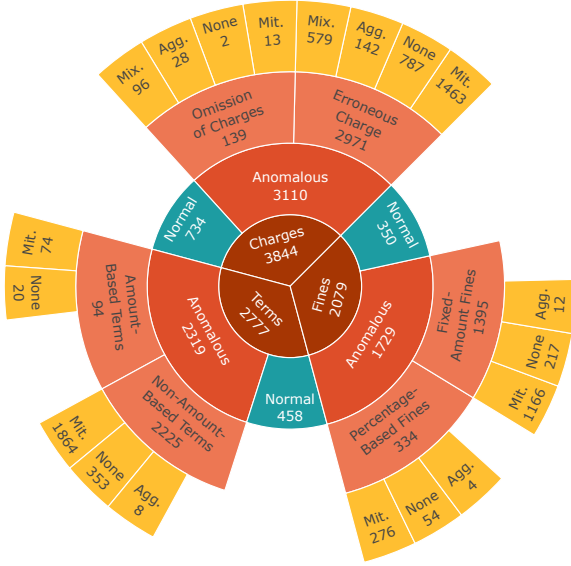


Figure 3: Distribution of AR-BENCH, highlighting normal versus anomalous cases and further breaking anomalies down by error type (e.g., omission, erroneous charges, fixed amounts, percentage-based fines) and sentencing factors (None, Mitigated, Aggravated, Mixed).

error correction task.

Following the automated synthesis pipeline, we conducted a two-stage manual quality evaluation. In the first stage, three legal experts first evaluated a stratified sample of 400 cases to ensure strict consistency between the reasoning processes and final judgments. Based on their review, we refined the injection pipeline by strictly enforce subject consistency and deleting residual text fragments. In the second stage, the experts performed a post-validation using a newly and independently sampled set of 400 cases. The final evaluation yielded a Fleiss’ κ of 0.79 and a pass rate of 91.5%.

3.5 Dataset Statistics

As illustrated in Figure 3, AR-BENCH comprises a total of 8,700 single-defendant criminal judgments. The dataset is comprehensively distributed across three core adjudication elements, yielding 3,844 cases for Charges, 2,777 for Prison Terms, and 2,079 for Fines. To effectively evaluate a model’s ability to avoid false positives (i.e., hallucinating errors in flawless documents), we maintained a specific ratio of normal to anomalous cases. For instance, within the Charges category, 3,110 judgments contain manually injected errors, while 734 remain in their original state. The mean character lengths for Case Facts, Reasoning, Judgment, and Law Articles are 380.9, 286.5, 87.6, and 557.9, re-

spectively (see Table 5 in the Appendix for detailed length distributions).

4 Experiments

The primary objective of this section is to systematically analyze current LLMs’ performance and limitations. Specifically, while our main evaluation (RQ1) assesses a wide range of mainstream LLMs, all subsequent analytical experiments (RQ2-RQ4) are conducted exclusively using Qwen3-Max (Yang et al., 2025a). We conduct experiments to address the following questions:

RQ1: How do existing LLMs perform on the three subtasks?

RQ2: How do different components in judgment documents and additional knowledge (such as legal articles) affect model performance?

RQ3: How do existing models perform on APPELLATE REVIEW across different error types?

RQ4: How do humans perform on the three subtasks of APPELLATE REVIEW?

4.1 Baselines

We categorize baselines into 3 groups: general-domain LLMs, reasoning-enhanced LLMs and domain-specific LLMs.

General-Domain LLMs. This group consists of 8 general LLMs, including GPT-5-2025-08-07 (no-thinking) (OpenAI, 2025), GPT-4o (Hurst et al., 2024), Qwen3-Max-2025-09-23 (Yang et al., 2025a), DeepSeek-V3.2-Chat (DeepSeek-AI et al., 2025), GLM-4.6-Chat (GLM et al., 2024), MiniMax-M2 (MiniMax, 2025), Qwen3-14B-Chat (Yang et al., 2025a), and Qwen3-8B-Chat (Yang et al., 2025a). These models are pre-trained on diverse, large-scale corpora to develop robust language understanding and generation capabilities.

Reasoning-Enhanced LLMs. This group contains a total of 6 reasoning LLMs, including GPT-5-2025-08-07 (thinking) (OpenAI, 2025), GLM-4.6-Thinking (GLM et al., 2024), Qwen3-Max-preview-Thinking (Yang et al., 2025a), DeepSeek-V3.2-Thinking (DeepSeek-AI et al., 2025), Kimi-K2-Thinking (Kimi et al., 2025), Qwen3-Next-80B-A3B-Thinking (Yang et al., 2025a). In contrast to general-domain LLMs that provide direct responses, these models are designed to generate internal Chain-of-Thought processes before arriving at a final answer. They have been optimized through Reinforcement Learning

Models	Error Detection				Error Classification				Error Correction			
	Acc	MaP	MaR	MaF1	Acc	MaP	MaR	MaF1	Acc	MaP	MaR	MaF1
GENERAL-DOMAIN LLMs												
GPT-5	86.40	78.63	85.67	81.14	87.69	89.32	87.57	88.16	58.09	56.26	50.02	50.44
GPT-4o	81.05	74.73	85.99	76.61	89.34	88.91	90.47	89.51	61.94	56.91	51.20	51.30
Qwen3-Max	92.03	86.57	89.78	88.03	94.75	95.01	94.05	94.48	74.13	<u>69.02</u>	65.49	<u>65.61</u>
DeepSeek-V3.2-Chat	78.87	71.54	80.47	73.26	84.67	86.71	82.04	83.88	65.68	61.55	58.86	58.04
GLM-4.6-Chat	87.72	80.23	86.66	82.68	91.82	92.45	90.35	91.25	70.99	67.16	60.66	61.48
MiniMax-M2	79.51	71.80	80.29	73.68	77.23	86.39	75.35	80.47	57.12	58.90	51.79	51.77
Qwen3-14B-Chat	88.05	80.66	87.25	83.16	85.41	87.72	84.55	84.78	55.97	57.69	46.84	46.70
Qwen3-8B-Chat	65.84	67.15	76.86	62.94	71.14	76.75	62.77	63.86	52.86	50.24	44.37	40.35
REASONING-ENHANCED LLMs												
GPT-5-Thinking	84.55	76.15	81.83	78.26	84.07	85.37	81.74	82.35	62.83	63.04	60.18	58.37
GLM-4.6-Thinking	90.34	84.11	87.22	85.52	91.20	90.18	90.47	90.29	68.86	64.99	61.79	61.48
Qwen3-Max-preview-Thinking	91.57	87.21	85.92	<u>86.54</u>	91.63	91.12	90.57	90.81	70.89	67.10	<u>66.43</u>	65.18
DeepSeek-V3.2-Thinking	87.40	80.03	<u>88.62</u>	82.89	88.51	90.28	87.03	88.62	<u>72.40</u>	69.89	68.18	67.36
Kimi-K2-Thinking	<u>91.70</u>	89.39	83.39	85.93	<u>92.55</u>	<u>93.31</u>	<u>91.85</u>	<u>92.57</u>	<u>71.55</u>	66.22	63.47	63.27
Qwen3-Next-80B-A3B-Thinking	89.19	<u>88.18</u>	75.88	79.99	88.66	88.02	86.64	87.27	63.51	56.82	56.56	54.55
DOMAIN-SPECIFIC LLMs												
Farui-Plus	48.05	60.16	63.89	47.27	66.22	63.43	64.21	63.56	52.60	52.27	43.08	41.10
LawLLM-7B	26.34	59.55	53.76	25.04	39.85	53.13	42.70	34.20	31.38	47.59	31.02	31.65

Table 2: Performance comparison of large language models on error detection, error classification, and error correction tasks, reported in terms of Accuracy (Acc), Macro-Precision (MaP), Macro-Recall (MaR), and Macro-F1 (MaF1) across general-domain, reasoning-enhanced, and domain-specific LLMs.

Settings	Error Detection		Error Classification		Error Correction	
	Acc	MaF1	Acc	MaF1	Acc	MaF1
S1	91.79	87.09	92.56	48.07	86.30	86.50
S2	97.06	95.09	99.97	49.99	97.71	96.01
S3	95.79	92.78	99.97	49.99	97.34	95.46
S4	94.63	91.47	96.67	49.15	95.46	90.93
S5	92.79	88.29	94.51	48.59	94.14	88.59

Table 3: Performance comparison of Qwen3-Max across 5 settings on three subtasks within the charge error subset, reported in terms of accuracy (Acc) and macro F1-score (MaF1).

to produce extended reasoning paths to enhance the precision of complex outputs.

Domain-Specific LLMs. This group of LLMs is composed of Farui-Plus (Alibaba Group, 2026) and LawLLM-7B (Yue et al., 2024b). Domain-specific large models are typically fine-tuned from general-domain LLMs, utilizing legal-domain corpora and task-optimized objectives to enhance the model’s capability in handling legal tasks.

4.2 Input Settings

We compare the impact of different input components in Section 4.4, and all other experiments use S4 unless otherwise specified.

S1: *Case facts + anomalous judgment* aligns with the legal judgment prediction paradigm.

S2: *S1 + normal reasoning process* contains sen-

tencing factors and cited law article numbers.

S3: *S2 + relevant law articles* provides law articles as additional legal knowledge.

S4: *Case facts + anomalous reasoning process + anomalous judgment* as the most realistic setting.

S5: *S4 + erroneously cited law articles* provides anomalous legal knowledge.

Specifically, normal reasoning process refers to the original reasoning process extracted from the original judgment. Anomalous reasoning process refers to the synthesized reasoning process created by modifying specific components of the original reasoning process to align with the injected erroneous judgment outcome.

4.3 Main Results (RQ1)

Table 2 reports the performance scores of all evaluated LLMs on three subtasks of AR-BENCH.

Performance across tasks reveals an interesting divergence: error classification consistently yields the highest scores, making it empirically the most tractable task for LLMs, despite error detection logically serving as its prerequisite. This phenomenon can be attributed to the conditional nature of the tasks and the drastically different target spaces. Our classification prompt explicitly informs the model that an error exists, reducing the task to merely categorizing a known anomaly into predefined classes (Charge, Prison Term, or Fine). In contrast, error



Figure 4: Performance comparison of Qwen3-Max across different error types using error detection, classification, and correction metrics, including Macro F1 (MaF1), ImpScore, and correction accuracy at 0.1 tolerance (Acc@0.1).

detection requires unguided skepticism to globally identify hidden flaws without any prior hints.

Expectedly, error correction remains the most challenging task, as it transitions from discriminative selection to generative precision. Furthermore, a domain-specific discrepancy emerges: specialized legal models significantly underperform general-domain and reasoning-enhanced LLMs. This indicates that existing domain models, primarily fine-tuned on flawless judicial templates and standard QA tasks, lack the adversarial diagnostic reasoning required for rigorous appellate review, severely degrading their performance.

4.4 Impact of Input Settings (RQ2)

Table 3 shows that introducing explicit reasoning substantially improves judgment review performance, as S2 and S4 consistently outperform the basic setting S1 across all subtasks. However, a *knowledge-as-noise* phenomenon emerges: incorporating relevant legal statutes provides no additional benefit and even degrades performance (S3 vs. S2, and S5 vs. S4). This suggests that LLMs still struggle with complex legal understanding in this context. We hypothesize that rather than selectively applying precise boundaries or exceptions like a human judge, the models might fail to prioritize the legally critical context within statutory provisions. Consequently, the inclusion of extensive legal text may act as a distractor, inadvertently introducing noise into the reasoning process.

4.5 Performance Across Error Types (RQ3)

The experimental results in Figure 4 reveal systematic limitations of existing LLMs in APPELLATE REVIEW. In correction tasks, models excel at identifying Erroneous Charges (E1) but struggle with

Subtask	Detection	Classification	Correction
Qwen3-Max	48.46	92.03	53.81
Human	49.15	93.64	65.66

Table 4: Human performance measured by Macro-F1 on a sampled subset of anomalous judgments for error detection, error classification, and error correction.

Omissions (E2). This gap implies that regardless of potential training data overlap, performance bottlenecks stem from holistic deductive reasoning rather than rote memorization of original cases. Furthermore, performance drops precipitously on fine prediction, particularly in percentage-based fine errors (E6), exposing the limitations of LLMs in reliably applying strict statutory formulas without external computational tools. Consequently, continuous evaluation metrics (ImpScore and Acc@0.1) provide a more fine-grained assessment of these arithmetic reasoning failures compared to interval-based classification, which may not fully capture meaningful boundary differences. A detailed case study can be found in Appendix B.4.

4.6 Human Performance Evaluation (RQ4)

We enlisted three certified legal experts to perform APPELLATE REVIEW on a stratified sample of 150 cases. Two experts evaluated the subset independently, with disagreements resolved via arbitration by a senior third expert. To ensure a fair comparison with the LLM, evaluators operated under identical context-only settings (S4) without internet search. As demonstrated in Table 4, human evaluators outperform the model overall, particularly on error correction, exposing a substantial gap between current LLM performance and expert-level legal diagnostic reasoning.

5 Conclusion

In this paper, we introduce AR-BENCH, a controlled benchmark designed to evaluate models' legal diagnostic reasoning underlying appellate review through error detection, classification and correction. Detailed analyses reveal the limitations of current LLMs, exposing a gap between general language generation and reliable legal diagnostic reasoning. AR-BENCH provides a new foundation for exploring judgment review tasks and improving legal diagnostic reasoning. We hope this resource catalyzes future efforts to build more robust legal AI. Future work will expand the range of error types and develop metrics for jointly evaluating multiple and cascading errors.

Limitations

Despite its contributions, this study has several limitations that should be acknowledged. The scope of judgments is limited to Chinese criminal cases and does not include Chinese civil cases nor common law precedents from jurisdictions such as England and the United States. This trade-off was made to concentrate resources and ensure data quality during the initial research phase. AR-BENCH does not include legal interpretation or argumentation tasks, primarily because high-quality legal reasoning annotations that rely heavily on experts are difficult to scale and lack fine-grained domain evaluation criteria. Relying on existing large language models for assessment would introduce model bias, compromising objectivity. AR-Bench intentionally focuses on legally verifiable correction targets, including charges, prison terms, and fines. These targets represent the core legal outcomes that must be determined before producing a complete appellate judgment document, while enabling objective automatic evaluation. Errors involving factual factors and potentially mixed fact-application factors fall outside the current scope and remain important directions for future research. We acknowledge these limitations and hope they will provide clear directions for future research.

Ethical Considerations

Our work may involve the following ethical issues: (1) Data Privacy and Offensive content. Our data is sourced from publicly available datasets, and sensitive information has been removed during the data construction process. To ensure strict data privacy

and comply with the Supreme People's Court guidelines for document publication, we implemented a rigorous anonymization pipeline for all datasets. Details can be found in Appendix A.2. The case facts in the public judgment shall not contain any legally unfounded or unnecessarily offensive or inappropriate content, provided that they comply with legal provisions and judicial ethics. (2) Data Bias. Publicly available documents on the China Judgments Online exhibit selectivity and may vary in disclosure across different regions and court levels. This could introduce biases in the dataset regarding case types, geographic distribution, and temporal span, thereby affecting the benchmark's universality. (3) Legality and Compliance of Data Use. The purpose of constructing this dataset is to explore the limits of existing models in the legal domain. Therefore, this dataset is intended solely for academic research and must not be used for actual judicial decision-making. (4) Potential Impact on Judicial Fairness. AI-assisted judgment review may help standardize adjudication criteria, but it also risks overlooking the unique circumstances of each case. As a benchmark, AR-BENCH does not provide any legal advice. Evaluation results from our datasets confirm that existing models still struggle to perform appellate review effectively. Given the complexity of legal cases, we emphasize that AI models should not replace human prosecutors or judges in the legal domain. We are committed to helping legal practitioners clarify the capabilities and boundaries of existing AI models, so they know when and how to use them.

Additionally, we used ChatGPT to polish the writing and are responsible for all the materials presented in this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U23B2056), in part by the Fundamental Research Funds for the Central Universities, and in part by the State Key Laboratory of Complex & Critical Software Environment.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoŧiuc-Pietro, and Vasileios Lampsos. 2016. [Predicting judicial decisions of the european court of human rights: A natural language processing perspective](#). *PeerJ computer science*, 2:e93.

- Alibaba Group. 2026. [Tongyi farui](#). Accessed: 2026-01-04.
- Zhenwei An, Quzhe Huang, Cong Jiang, Yansong Feng, and Dongyan Zhao. 2022. [Do charge prediction models learn legal theory?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3757–3768, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jianjun Bai. 2006. [An empirical study on the change of accusation](#). *Chinese Journal of Law*, (04):51–62.
- Sheng Bi, Zhiyao Zhou, Lu Pan, and Guilin Qi. 2023. [Judicial knowledge-enhanced magnitude-aware reasoning for numerical legal judgment prediction](#). *Artificial Intelligence and Law*, 31(4):773–806.
- Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. [Neural legal judgment prediction in English](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. 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](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367, Hong Kong, China. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2026. [LEXam: Benchmarking legal reasoning on 340 law exams](#). In *The Fourteenth International Conference on Learning Representations*.
- 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.
- Mariano Felice and Ted Briscoe. 2015. [Towards a standard evaluation method for grammatical error detection and correction](#). In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 578–587, Denver, Colorado. Association for Computational Linguistics.
- Yi Feng, Chuanyi Li, and Vincent Ng. 2022. [Legal judgment prediction via event extraction with constraints](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664, Dublin, Ireland. Association for Computational Linguistics.
- Leilei Gan, Baokui Li, Kun Kuang, Yating Zhang, Lei Wang, Anh Luu, Yi Yang, and Fei Wu. 2023. [Exploiting contrastive learning and numerical evidence for confusing legal judgment prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12174–12185, Singapore. Association for Computational Linguistics.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. [Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.
- Carissa Byrne Hessick and F Andrew Hessick. 2008. [Appellate review of sentencing decisions](#). *Ala. L. Rev.*, 60:1.
- Sophie Hoquet-McKee. 1981. [The french supreme court: The cour de cassation](#). *City London L. Rev.*, page 1.
- Changming Hu. 2024. [Solicitation behavior and corruption crimes](#). *Chengchi Law Review*, 178(178):1–35.
- 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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- 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](#). *Preprint*, arXiv:2505.16514.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow,

- Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. [A multi-task benchmark for korean legal language understanding and judgement prediction](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 32537–32551. Curran Associates, Inc.
- Team Kimi, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, and 150 others. 2025. [Kimi k2: Open agentic intelligence](#). *Preprint*, arXiv:2507.20534.
- Reed C Lawlor. 1963. [What computers can do: analysis and prediction of judicial decisions](#). *American Bar Association Journal*, pages 337–344.
- Ang Li, Qiangchao Chen, Yiquan Wu, Xiang Zhou, Kun Kuang, Fei Wu, and Ming Cai. 2024a. [From graph to word bag: Introducing domain knowledge to confusing charge prediction](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7469–7479, Torino, Italia. ELRA and ICCL.
- Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, Wuyue Wang, Yiqun Liu, and Minlie Huang. 2025. [LegalAgentBench: Evaluating LLM agents in legal domain](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (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. 2024b. [Lexeval: A comprehensive chinese legal benchmark for evaluating large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 25061–25094. Curran Associates, Inc.
- Michael Lindemann and Fabienne Lienau. 2020. [Mechanisms for correcting judicial errors in germany](#). *Erasmus L. Rev.*, 13:73.
- Bulou Liu, Zhenhao Zhu, Qingyao Ai, Yiqun Liu, and Yueyue Wu. 2024. [Ledqa: A chinese legal case document-based question answering dataset](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, page 5385–5389, New York, NY, USA. Association for Computing Machinery.
- Chao-Lin Liu and Chwen-Dar Hsieh. 2006. [Exploring phrase-based classification of judicial documents for criminal charges in chinese](#). In *Foundations of Intelligent Systems*, pages 681–690, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Dugang Liu, Weihao Du, Lei Li, Weike Pan, and Zhong Ming. 2022. [Augmenting legal judgment prediction with contrastive case relations](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2658–2667, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. [Automatic judgment prediction via legal reading comprehension](#). In *Chinese Computational Linguistics*, pages 558–572, Cham. Springer International Publishing.
- MiniMax. 2025. [Minimax m2 & agent: Ingenious in simplicity](#). Technical report, MiniMax. Accessed: 2026-01-04.
- Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. [Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark](#). In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill A. H. Chevalley, Claude Humbel, Christophe Gösen, Lorenzo Tanzi, Thomas Lüthi, Stefan Palombo, Spencer Poff, Boling Yang, Nan Wu, Matthew Guillod, Robin Mamié, Daniel Brunner, Julio Pereyra, and Niko Grupen. 2025. [SwiLTra-bench: The Swiss legal translation benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14894–14916, Vienna, Austria. Association for Computational Linguistics.
- OpenAI. 2025. [Introducing GPT-5](#). Blog Post. Accessed: 2026-2-18.
- Lester B Orfield. 1936. [History of criminal appeal in england](#). *Mo. L. Rev.*, 1:326.
- Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. [CaseSummarizer: A system for automated summarization of legal texts](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 258–262, Osaka, Japan. The COLING 2016 Organizing Committee.
- PRC. 2025. [Judicial statistics in China](#).
- Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. [Averitec: A dataset for real-world claim verification with evidence from the web](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167. Curran Associates, Inc.

- Jaehyung Seo, Dahyun Jung, Jaewook Lee, Yongchan Chun, Dongjun Kim, Hwijung Ryu, Donghoon Shin, and Heuseok Lim. 2025. [KoLEG: On-the-fly Korean legal knowledge editing with continuous retrieval](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9191–9217, Suzhou, China. Association for Computational Linguistics.
- Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. [Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 13158–13173. Curran Associates, Inc.
- Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. [A comprehensive evaluation of large language models on legal judgment prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore. Association for Computational Linguistics.
- Weihang Su, Baoqing Yue, Qingyao Ai, Yiran Hu, Jiaqi Li, Changyue Wang, Kaiyuan Zhang, Yueyue Wu, and Yiqun Liu. 2025. [Judge: Benchmarking judgment document generation for chinese legal system](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25*, page 3573–3583, New York, NY, USA. Association for Computing Machinery.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Santosh T.y.s.s., Mohamed Hesham Elganayni, Stanisław Sójka, and Matthias Grabmair. 2024. [Incorporating precedents for legal judgement prediction on European court of human rights cases](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3743–3750, Miami, Florida, USA. Association for Computational Linguistics.
- S Sidney Ulmer. 1963. [Quantitative analysis of judicial processes: Some practical and theoretical applications](#). *Law and Contemporary Problems*, 28(1):164–184.
- Rohit Upadhyia and Santosh T.y.s.s. 2025. [LexCLiPR: Cross-lingual paragraph retrieval from legal judgments](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13971–13993, Vienna, Austria. Association for Computational Linguistics.
- J Clifford Wallace. 2005. [Improving the appellate process worldwide through maximizing judicial resources](#). *Vand. J. Transnat'l L.*, 38:187.
- Margaret YK Woo. 1989. [The right to a criminal appeal in the people's republic of china](#). *Yale J. Int'l L.*, 14:118.
- Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. [Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075, Singapore. Association for Computational Linguistics.
- Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. [Cail2018: A large-scale legal dataset for judgment prediction](#). *Preprint*, arXiv:1807.02478.
- Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. [Distinguish confusing law articles for legal judgment prediction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095, Online. Association for Computational Linguistics.
- Nuo Xu, Pinghui Wang, Junzhou Zhao, Feiyang Sun, Lin Lan, Jing Tao, Li Pan, and Xiaohong Guan. 2024. [Distinguish confusion in legal judgment prediction via revised relation knowledge](#). *ACM Trans. Inf. Syst.*, 43(1).
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Hongkun Yang, Lionel Z. Wang, Wei Fan, Yiran Hu, Lixu Wang, Chenyu Liu, Yu Zeng, Shenghong Fu, Lei Gong, Zhengxin Zhang, Haoyang Li, Jiexin Zheng, and Xin Xu. 2026. [Appellategen: A benchmark for appellate legal judgment generation](#). *Preprint*, arXiv:2601.01331.
- Tiankai Yang, Yi Nian, Li Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan A. Rossi, Kaize Ding, Xia Hu, and Yue Zhao. 2025b. [AD-LLM: Benchmarking large language models for anomaly detection](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1524–1547, Vienna, Austria. Association for Computational Linguistics.
- Linan Yue, Qi Liu, Binbin Jin, Han Wu, and Yanqing An. 2024a. [A circumstance-aware neural framework for explainable legal judgment prediction](#). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5453–5467.
- Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. [Neurjudge: A circumstance-aware](#)

- neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 973–982, New York, NY, USA. Association for Computing Machinery.
- Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2024b. [Lawllm: Intelligent legal system with legal reasoning and verifiable retrieval](#). In *Database Systems for Advanced Applications*, pages 304–321, Singapore. Springer Nature Singapore.
- Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. [Contrastive learning for legal judgment prediction](#). *ACM Trans. Inf. Syst.*, 41(4).
- Yuan Zhang, Wanhong Huang, Yi Feng, Chuanyi Li, Zhiwei Fei, Jidong Ge, Bin Luo, and Vincent Ng. 2024. [LJPCheck: Functional tests for legal judgment prediction](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5878–5894, Bangkok, Thailand. Association for Computational Linguistics.
- Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018a. [Legal judgment prediction via topological learning](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.
- Haoxi Zhong, Chaojun Xiao, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018b. [Overview of cail2018: Legal judgment prediction competition](#). *Preprint*, arXiv:1810.05851.
- Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. [Jecqa: A legal-domain question answering dataset](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9701–9708.
- Weimin Zuo. 2018. [Response to litigation explosion in china:based on empirical analysis of the trial practice of the court in w district in the past thirty years](#). *Chinese Legal Science*.

A Additional Details of AR-BENCH

A.1 License and Copyright Issues

AR-BENCH is derived from the JuDGE dataset². The original JuDGE project was released under the MIT License in its official repository. In strict compliance with the permissive terms of this license, we adapted the dataset by (1) collecting the raw judicial documents, (2) constructing annotations tailored for the APPELLATE REVIEW task, and (3) optimizing the data format for model evaluation. To foster open research, AR-BENCH is also released under the MIT License, retaining original copyright notices without imposing any additional restrictions.

A.2 Anonymization

The raw legal documents were obtained from China Judgments Online³. By law, these public documents are mandated to undergo strict de-identification prior to publication according to the Provisions of the Supreme People’s Court (e.g., removing specific names, IDs, and contact details). However, rather than solely relying on upstream institutional compliance, we implemented an additional, rigorous anonymization pipeline to mitigate any potential oversights. During our data processing and expert annotation phases, we utilized regular expressions and manual spot-checks to actively screen for, detect, and mask any residual personally identifiable information (PII) or sensitive geographic details. Consequently, rigorous efforts were made to minimize the risk of personal data disclosure.

A.3 Dataset Statistics

To provide a more comprehensive understanding of the textual characteristics within AR-BENCH, we detail the character-level length distributions of the four core document components in Table 5. Notably, the Law Articles and Fact Descriptions exhibit the highest average lengths (557.9 and 380.9 characters, respectively), reflecting the extensive legal context and factual details required for diagnostic reasoning. Conversely, the final Judgments are highly concise, averaging only 87.6 characters. Furthermore, the substantial variance between the maximum and minimum lengths across all components highlights the diverse range of case complexities

encapsulated within our benchmark, demanding robust long-context processing capabilities from the evaluated LLMs.

Component	Max Length	Min Length	Mean Length
Fact Description	1692.0	89.0	380.9
Reasoning Process	2567.0	97.0	286.5
Judgment	1027.0	19.0	87.6
Law Article	1622.0	39.0	557.9

Table 5: Character-level length distribution of sample components in our AR-BENCH.

A.4 Data Annotation and Anomaly Synthesis

The initial dataset annotation and anomaly injection were conducted by three researchers with foundational legal knowledge. To ensure professional rigor, a certified procurator was invited to train the annotators on the principles of judgment errors and judicial logic prior to the annotation process. All annotators strictly adhered to unified operational guidelines.

Due to space constraints, we present representative examples of the synthesized anomalous judgments (e.g., Erroneous Conviction and Omission of Charges) in Table 6 and Table 7. The complete set of examples across all six error types (E1–E6) is provided in our supplementary materials.

During the dataset construction process, we engaged three legal experts to assist with quantitative quality control. All experts have passed the National Uniform Legal Profession Qualification Examination and possess specialized expertise in criminal law. They were compensated at a professional hourly rate exceeding the local minimum wage. These experts conducted sampling evaluations for both the label extraction and anomaly synthesis stages. Quality Evaluation Guidelines are provided below:

Objective: In the following tasks, we will provide you with real-world court judgments, their extracted annotations, and synthesized erroneous judgments generated through our modifications. Based on your professional legal knowledge, please independently assess: (1) the accuracy of the extracted annotations, and (2) the logical coherence of the synthetic judgments. For each question, please provide two binary labels based on the criteria below. Data from this assessment is for academic research only.

²<https://github.com/oneal2000/JuDGE>

³<https://wenshu.court.gov.cn/>

1. Annotation Accuracy Label

Determine whether the charges, prison terms, fines, and legal articles extracted in the original judgment annotations are completely accurate.

- **1 (Correct):** The extracted annotations are completely accurate.
- **2 (Mostly Correct):** The extracted annotations are generally accurate, but the extracted legal article numbers contain errors.
- **0 (Incorrect):** The extracted annotations contain obvious errors compared to the original judgment (e.g., incorrect charges, fines, or prison terms).

2. Synthetic Judgment Logical Consistency Label

Determine whether the synthetic judgment remains logically consistent and does not leak the original charges, fines, and prison terms.

- **1 (Yes):** No collateral errors were introduced; the text modifications are exceptionally clean.
- **0 (No):** There are obvious logical errors (e.g., contextual inconsistencies that lead to the leakage of the ground-truth labels).

A.5 Human Performance Evaluation

To quantify the capability gap between large language models and human professionals, we conducted a human baseline evaluation. To maintain independence and objectivity, we recruited three additional independent legal experts to perform the baseline testing. They have all passed the National Uniform Legal Profession Qualification Examination and are familiar with criminal law. They were compensated at an hourly rate exceeding the local minimum wage. Evaluators were provided with the same input settings (S4) as the models to assess a randomly sampled subset of 150 cases. Under a controlled environment without internet access, they were required to sequentially complete the three core subtasks: Error Detection, Error Classification, and Error Correction.

B Additional Details on Experiments

B.1 Hyperparameter settings

We report the hyperparameters used for the LLMs in Table 8. When evaluating LLMs, we set the temperature to 0 to minimize the variance introduced

by random sampling. We employed the same hyperparameter settings as in the original papers when training SLMs on the CAIL2018 dataset.

B.2 Computational Budget

All experiments were conducted on an Ubuntu server equipped with a 32-core Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz and 1 Tesla V100-PCIE-32GB. Additionally, we fine-tuned legal judgment prediction models to evaluate their performance. The CUDA version was 13.0, the Python version was 3.10, the PyTorch version was 2.4.0+cu121, and the tensorflow-gpu version was 1.14.0. The GPU hours of training are shown in Table 9.

B.3 Comparison with small language model

We further conducted experiments to address the following questions: Can legal judgment prediction models be applied to the error correction subtask? How do they perform on existing datasets and AR-BENCH? We fine-tuned a group of small language models (SLMs) including TopJudge (Zhong et al., 2018a), LADAN (Xu et al., 2020), and NeurJudge (Yue et al., 2021). These models are SOTA pre-trained models fine-tuned for legal judgment prediction. We use charge error correction to assess their generalization performance on our dataset.

Table 10 shows that on our AR-BENCH, most existing SOTA small language models achieve higher accuracy but perform worse on other metrics. Compared with CAIL, AR-BENCH places greater emphasis on error detection capability. Cases that are prone to misjudgment often involve low-frequency crimes with limited samples, which are inherently more difficult to learn than common charges. Therefore, Macro-F1 (MaF1) is a more meaningful evaluation metric in this setting. The relatively low MaF1 scores indicate both the challenging nature of our dataset and the limitations of current models in handling rare and error-prone cases.

B.4 Case Study

In this section, we analyze the limitations of the evaluated LLMs through several representative cases. Specifically, we select the best-performing model, Qwen3-Max, and conduct an in-depth analysis of its failure cases. We summarize three categories of failure cases, as illustrated below:

Breakdowns in Concept Mapping and Holistic Deduction

- **Failures to distinguish normative concurrence:** When the conduct in a case gives rise to normative concurrence between multiple offenses (the crime of fraud versus the crime of contract fraud), the model fails to apply the principle that special laws (i.e., the crime of contract fraud) take precedence over general laws (i.e., the crime of fraud), and erroneously treats this broad misclassification as correct, as shown in Table 11.
- **Limited ability to recognize complex sentencing factors:** When a defendant commits a new crime while on probation, the offenses should be consolidated for sentencing; however, the model fails to capture key information in the reasoning process, leading it to misclassify a correct judgment as incorrect, as shown in Table 12.
- **Deficiencies in the Identification of Omission-Type Anomalies:** When constructing erroneous judgments exhibiting Omission of Charges (E2), we retained only one charge along with its corresponding reasoning, while removing all segments related to the other charges. The model failed to identify this as an Omission of Charges, as shown in Table 13.

Failures in Numerical Sensitivity and Rule Boundaries

- **Failure to map statutory numerical thresholds to sentencing tiers:** The model is unable to map the amounts involved in the case to the correct sentencing range based on the facts of the case, making it impossible to identify sentencing errors, as shown in Table 14.
- **Misalignment of fine-grained scales in sentencing correction:** For economic crimes and other offenses with broad sentencing ranges, the model can identify the correct direction of adjustment (aggravation or mitigation), but lacks an accurate understanding of the severity of the case, leading to deviations in the quantitative calculations, as shown in Table 15.

Contextual Distraction and Diagnostic Misattribution

- **Misattribution of root causes in cascading errors:** In cases of cascading erroneous judgments, the root error leads to errors in both the sentence

and the fine, making such cases particularly challenging. Models tend to attribute these errors to superficial issues (e.g., sentencing error) without recognizing the underlying cause (e.g., charge error), as shown in Table 16.

- **Anchoring bias:** Although we explicitly instructed the model in the prompt that there is a charge error in the judgment and asked it to correct the incorrect charge, the model still repeats the incorrect charge from the erroneous judgment, as shown in Table 17.
- **Diagnostic logic misalignment in long-tail charges:** When faced with rare criminal charges that have long names and complex characteristics (e.g., “the crime of illegally selling invoices used to fraudulently claim export tax rebates or offset tax liabilities”), the model’s semantic representation tends to classify them as errors in the charge description rather than analyzing the actual errors in the prescribed sentence, as shown in Table 18.

B.5 Prompts

To ensure reproducibility and prevent prompt engineering biases, we designed structured, zero-shot prompt templates for all LLM evaluations. Each prompt strictly consists of four modular components: Role and Mission, Core Task, Analytical Basis, and a zero-tolerance Output Format.

Due to space constraints, we present the foundational prompt template for the Error Detection task in Prompt B.6.2. The prompts for Error Classification and Error Correction follow the same structural paradigm, with only minor task-specific modifications applied to the target variables and JSON schemas. The complete suite of prompts for all subtasks is provided in our supplementary materials and repository.

B.6 Pipeline Reliability and Instruction-Following Failures

B.6.1 Strict Zero-Exclusion Policy

To address concerns regarding the end-to-end reliability of large language models in real-world judicial pipelines, we implement a strict zero-exclusion policy for all evaluations. In practical applications, models must output results in valid JSON strings for downstream automated processing. Therefore, if a model fails to follow the prompt’s structural constraints, the prediction is directly recorded as incorrect. This approach explicitly avoids masking

pipeline reliability issues and ensures our metrics reflect actual deployment utility rather than overly optimistic theoretical capability.

B.6.2 Discussion on Instruction-Following and Extraction Bias

Under this strict evaluation paradigm, we observed that while some reasoning-enhanced LLMs demonstrated strong internal legal logic, their metrics occasionally suffered due to formatting breakdowns (see Table 19). These models tend to over-explain their reasoning, prioritizing conversational completeness over strict schema adherence. Crucially, we deliberately avoided using regular expressions to forcefully extract labels from these conversational outputs. We observed that when models break the JSON format, it is frequently a manifestation of "hedging" or uncertainty regarding complex legal logic. Forcefully extracting a specific label from a convoluted explanation introduces significant subjective evaluation bias. This highlights a critical finding: reliable legal AI requires not only accurate diagnostic reasoning but also robust format adherence to integrate successfully into automated review systems.

Fact	Upon trial, it was ascertained that: 1. On April 27, 2008, the defendant XX, without obtaining the right to use three rooms of land at the East Sanjianfang base in Ying County, collected a deposit of RMB 8,600 from the victim Wang XX under the pretext of selling the three-room residential land to him. In the spring of 2010, when the victim Wang XX prepared to build a house on the land, he discovered that the land had already been built upon by others and was not owned by the defendant XX, which the defendant later admitted. 2. In early November 2009, Wang XX1 entrusted the defendant XX to use his residential land use certificate as collateral to borrow RMB 15,000 from others. On November 6 of the same year, the defendant XX used the land use certificate of the house as collateral to borrow RMB 60,000 from the victim Yuan XX, and handed RMB 15,000 to Wang XX1. On January 1, 2010, the defendant XX borrowed an additional RMB 10,000 from the victim Yuan XX, later changing his contact information. The victim Yuan XX filed a civil suit to claim rights against XX, which was tried in absentia, and in 2012 entered enforcement proceedings. 3. In 2009, Yang XX, through the defendant XX, borrowed RMB 70,000 from the victim Shi XX using a relocation housing contract, and later borrowed an additional RMB 40,000 from Shi XX. During this period, Yang XX repaid RMB 83,000 to the victim Shi XX, leaving RMB 27,000 used by the defendant XX. The defendant XX issued an IOU for the remaining RMB 27,000 and agreed to repay it in full by January 30, 2010. Later, the defendant XX left Ying County and changed his contact information.	
Reasoning	Original Judgment Document	Anomalous Judgment Document
	Upon review, the Court finds that the defendant XX with the purpose of illegal possession, used methods of concealing the truth to defraud others of property, involving a huge amount, and that his conduct constitutes the crime of contract fraud Therefore, in accordance with Articles 244 and 64 of the Criminal Law of the People's Republic of China, the judgment is rendered as follows:	Upon review, the Court finds that the defendant XX with the purpose of illegal possession, used methods of concealing the truth to defraud others of property, involving a huge amount, and that his conduct constitutes the crime of fraud Therefore, in accordance with Articles 25, 53, 64, 266, and 52 of the Criminal Law of the People's Republic of China, the judgment is rendered as follows:
Judgment	Original Judgment Document	Anomalous Judgment Document
	I. The defendant XX is convicted of the crime of contract fraud and is sentenced to fixed-term imprisonment of three years and six months , and a fine of RMB 5,000 ...	I. The defendant XX is convicted of the crime of fraud and is sentenced to fixed-term imprisonment of three years and nine months , and a fine of RMB 80,000 ...
Prison Term	Original Judgment Document	Anomalous Judgment Document
	fixed-term imprisonment of three years and six months	fixed-term imprisonment of three years and nine months
Fine	Original Judgment Document	Anomalous Judgment Document
	a fine of RMB 5000	a fine of RMB 80000
Charge	Original Judgment Document	Anomalous Judgment Document
	contract fraud	fraud
Law Articles	Original Judgment Document	Anomalous Judgment Document
	"244", "64"	"25", "53", "64", "266", "52"
Factor	None	

Table 6: Example of Erroneous Conviction.

Fact	Upon trial, it was ascertained that on March 13, 2018, at 23:00, the defendant XX, while driving a taxi under the influence of alcohol, drove in the opposite direction on the road in front of the north gate of Yiju Mediterranean in Kangping County and collided with a normally driven car with license plate number Zhe E8T792. Traffic police officers XX and XX of the Kangping County Public Security Bureau arrived at the scene to handle the accident. When the defendant XX was brought to the outpatient department of Kangping County People's Hospital for blood collection, he kicked Officer XX with his feet. After forced blood collection, he punched XX in the face and verbally abused him, and was later subdued. The accident was caused by XX's drunk driving, and he was found fully responsible for the accident. Forensic examination detected ethanol in the defendant XX's venous blood at a concentration of 231.7 mg/100 ml.	
Reasoning	Original Judgment Document	Anomalous Judgment Document
	The Court holds that the defendant... committed the crimes of obstruction of official duties and dangerous driving . The facts of the crimes are clear, and the evidence is reliable and sufficient. The defendant shall be sentenced for obstruction of official duties and dangerous driving, to be combined for multiple crimes. In accordance with Articles 277, 133-1, 65(1), 67(3), 42, 44, 45, 47, 69(2), and 52 of the Criminal Law of the People's Republic of China, the judgment is rendered as follows:	The Court holds that the defendant's conduct constitutes the crime of dangerous driving . The public prosecution's allegations are upheld. In accordance with Articles 133-1(1), 67(1), 42, 44, 52, 77(1), 69, 45, and 47 of the Criminal Law of the People's Republic of China, the judgment is rendered as follows:
Judgment	Original Judgment Document	Anomalous Judgment Document
	The defendant XX is convicted of the crime of obstruction of official duties and is sentenced to fixed-term imprisonment of ten months; the defendant XX is convicted of the crime of dangerous driving and is sentenced to criminal detention of three months and fined RMB 3,000 . The sentences are combined, and the final execution is fixed-term imprisonment of ten months and a fine of RMB 3,000 (already paid) ...	The defendant is convicted of the crime of dangerous driving and sentenced to criminal detention of two months and fined RMB 2,000 , which, after revocation of probation and combination of sentences with fixed-term imprisonment of eleven months , results in the execution of fixed-term imprisonment of eleven months and a fine of RMB 2,000 ...
Prison Term	Original Judgment Document	Anomalous Judgment Document
	fixed-term imprisonment of ten months	fixed-term imprisonment of eleven months
Fine	Original Judgment Document	Anomalous Judgment Document
	a fine of RMB 3,000	a fine of RMB 2,000
Charge	Original Judgment Document	Anomalous Judgment Document
	"obstruction of official duties", "dangerous driving"	"dangerous driving"
Law Articles	Original Judgment Document	Anomalous Judgment Document
	"69", "67", "44", "47", "42", "133", "45", "277", "52"	"69", "67", "47", "77", "133", "42", "45", "44", "52"
Factor	All	

Table 7: Example of Omission of Charges.

Hyperparameter	G-LLMs	R-LLMs	D-LLMs
frequency_penalty	0	0	0
logprobs	false	false	false
presence_penalty	0	0	0
temperature	0	0	0
max_output_tokens	2048	2048	1024

Table 8: Hyperparameter settings. G-LLMs is short for General-Domain LLMs, R-LLMs is short for Reasoning-Enhanced LLMs, and D-LLMs is short for Domain-Specific LLMs.

Model	Training Set	GPU hours
LADAN+MIL	CAIL small	1
TopJudge	CAIL small	0.6
NeurJudge	CAIL small	4
LADAN+MIL	CAIL big	29
TopJudge	CAIL big	17
NeurJudge	CAIL big	98

Table 9: Computational Budget

Models	Training Set	Test Set	Charges				Prison Terms			
			Acc	MaP	MaR	MaF1	Acc	MaP	MaR	MaF1
TopJudge	CAIL-S	CAIL-S	71.61	66.83	73.43	67.28	34.65	27.19	30.91	26.47
		AR-BENCH	90.32	46.83	45.73	43.41	62.30	18.70	23.27	18.20
	CAIL-B	CAIL-B	85.49	33.40	36.60	32.18	49.77	26.24	29.62	26.57
		AR-BENCH	91.60	26.75	36.54	29.64	63.67	20.57	20.40	18.19
LADAN	CAIL-S	CAIL-S	84.67	82.58	82.35	81.93	35.57	31.25	32.58	31.76
		AR-BENCH	93.72	55.75	51.65	52.36	63.33	23.90	26.07	23.44
	CAIL-B	CAIL-B	96.01	82.00	86.05	83.36	58.77	43.60	49.63	45.19
		AR-BENCH	95.66	49.38	57.95	52.89	67.45	33.49	37.76	33.46
NeurJudge	CAIL-S	CAIL-S	83.06	81.04	81.41	80.60	37.79	32.11	35.79	33.56
		AR-BENCH	90.23	85.41	84.43	84.30	64.34	35.57	30.58	28.23
	CAIL-B	CAIL-B	89.07	55.08	71.44	58.41	50.58	24.52	39.60	26.39
		AR-BENCH	89.56	79.78	85.44	81.25	63.14	22.51	28.21	21.59

Table 10: Performance comparison of TopJudge, LADAN, and NeurJudge trained on CAIL small (CAIL-S) and CAIL big (CAIL-B), evaluated on CAIL and AR-BENCH test sets, reporting Accuracy (Acc), Macro Precision (MaP), Macro Recall (MaR), and Macro F1 (MaF1) for charge prediction and prison term prediction.

Key Fact Snippet	... In 2009, Yang XX borrowed 70,000 RMB from the victim Shi XX through the defendant Hou XX using a resettlement housing contract; subsequently, Yang XX borrowed an additional 40,000 RMB from Shi XX in installments...
Erroneous Judgment	The defendant Hou XX was convicted of fraud and sentenced to three years and nine months in prison, along with a fine of 80,000 RMB.
Model Output	The judgment is correct.
Ground Truth	The judgment is incorrect.
Error Analysis	The correct charge is the crime of contract fraud, not the crime of fraud. The model failed to map the criminal characteristics to the offense of contract fraud based on the key information in the case.

Table 11: A failure case in distinguishing normative concurrence.

Key Reasoning Snippet	... The defendant, Song XX, committed a new crime while on probation; therefore, the probation should be revoked, and the sentences for the multiple offenses should be combined...
Judgment	The defendant Song XX was found guilty of credit card fraud and sentenced to five months of detention and a fine of 20,000 RMB. Combined with Criminal Judgment No. (2015) Chang Xing Chu Zi No. 152, which previously sentenced him to one year of fixed-term imprisonment, suspended for one year, and a fine of 20,000 RMB for the crime of selling exit and entry documents, the suspension is hereby revoked. A combined sentence of one year of fixed-term imprisonment, suspended for one year, and a total fine of 40,000 RMB is ordered for enforcement.
Model Output	The judgment is incorrect.
Ground Truth	The judgment is correct.
Error Analysis	The defendant had previously been convicted of the crime of selling exit and entry documents and committed a new offense while on probation (as explicitly stated in the reasoning section); therefore, the sentences should have been imposed concurrently. However, the model determined that the judgment was erroneous when handling such complex sentencing factors.

Table 12: A failure case showing the model’s limited ability to recognize complex sentencing factors.

Key Fact Snippet	... At approximately 11:00 a.m. on May 23, 2019, the defendant Li XX was apprehended by police officers at his home. The officers seized one small packet of suspected methamphetamine (net weight: 4.76 grams) and one small packet of methamphetamine tablets (net weight: 1.84 grams) from Li XX’s person, as well as one small packet of suspected methamphetamine (net weight: 9.36 grams) and two small packets of methamphetamine tablets (net weights of 1.13 grams and 0.57 grams, respectively) from a drawer in the computer desk in the defendant’s bedroom. ... On the morning of May 16, 2019, the afternoon of May 18, 2019, the morning of May 20, 2019, and the evening of May 22, 2019, at his residence in XX, Beibei District, Chongqing, the defendant Li XX harbored a drug user, Liu XX, and they jointly consumed methamphetamine-type drugs by heating and inhaling them. The defendant Li XX provided a venue for others to use drugs on four occasions within one year...
Erroneous Judgment	The defendant was found guilty of illegal possession of narcotics and sentenced to fourteen years in prison, deprived of political rights for one year, and fined 100,000 yuan.
Model Output	The defendant should be convicted of illegal possession of drugs.
Ground Truth	The defendant should be convicted of the crimes of illegal possession of drugs and harboring others for the purpose of drug use.
Error Analysis	The model is unable to map multiple key pieces of information from the case facts (illegal possession of drugs and harboring others to use drugs) to their respective charges. Consequently, it fails to identify and correct the charges omitted in the erroneous judgment.

Table 13: A failure case in identifying omission-type anomalies.

Key Fact Snippet	In May 2013, the defendant Liu XX purchased low-grade “Tianxiaxiu” brand cigarettes and produced mid-range “Yunyan” and “Yuxi” brand cigarettes by replacing the filters. By September 5, 2013, when the case was uncovered, the defendant Liu XX and others had produced a total of 1,024 cartons of finished cigarettes. Of these, 967.3 cartons were sold at a price of 62 RMB per carton, generating a sales revenue of 59,972.6 RMB. ... According to an appraisal by the Sichuan Provincial Tobacco Quality Supervision and Testing Station, the seized cigarettes were all counterfeit and substandard...
Erroneous Judgment	The defendant, Liu XX, was convicted of the crime of producing and selling counterfeit or substandard products and sentenced to three years and one month in prison, along with a fine of 50,000 RMB.
Model Output	The judgment is correct.
Ground Truth	The judgment is incorrect.
Error Analysis	The model is unable to map the sales amount in the case facts to the correct sentencing range, making it impossible to identify the error in the judgment. According to Article 140 of the Criminal Law: Any producer or seller who mixes up or adulterates products, passes fake imitations for genuine, sells seconds at best quality price, or passes unqualified products as qualified ones, with a sales amount of not less than 50,000 RMB and not more than 200,000 RMB, is to be sentenced to not more than two years of fixed-term imprisonment or criminal detention and may in addition or exclusively be sentenced to a fine of not less than 50 percent and not more than 200 percent of the sale amount.

Table 14: A failure case in mapping statutory numerical thresholds to sentencing tiers.

Key Fact Snippet	On November 25, 2016, the defendants Wang XX, Fang XX, and Wu XX were placed under administrative detention for drug use. During their detention, Wu XX wrote down the login and payment passwords for his Alipay and WeChat accounts on a piece of cardboard from a soap box and passed it through the window of the detention cell door to Wang XX, who was taking a break outside. After defendant Wang XX was released upon completion of his administrative detention, he logged into Wu XX’s Alipay and WeChat accounts via his mobile phone and transferred funds—proceeds from Wu XX’s drug trafficking—from those accounts into his own account... In summary, Wang XX transferred 133 RMB by logging into Wu XX’s Alipay and WeChat accounts.
Erroneous Judgment	The defendant, Wang XX, was convicted of money laundering and sentenced to six years and three months in prison, along with a fine of twenty RMB.
Model Output	The defendant should be sentenced to five years in prison.
Ground Truth	The defendant should be sentenced to four months of detention.
Error Analysis	The model fails to accurately assess the severity of the offense based on the amount involved, resulting in a significant deviation from the appropriate sentence. According to Article 191 of the Criminal Law, an offender convicted of money laundering shall be sentenced to fixed-term imprisonment of not more than five years or detention.

Table 15: A failure case showing the misalignment of fine-grained scales in sentencing correction.

Key Fact Snippet	The trial established that at 11:00 p.m. on March 13, 2018, the defendant Gao XX, while driving a taxi under the influence of alcohol, was traveling in the wrong direction on the road ... when he collided with a sedan ... that was traveling in the correct direction. ...Police officers Zhang XX, Wang XX, and others arrived at the scene to handle the incident...the defendant Gao XX kicked traffic police officer Wang XX. After the blood sample was forcibly collected, he punched Zhang XX in the face and hurled insults, before being subdued...
Erroneous Judgment	The defendant was found guilty of dangerous driving and sentenced to two months of detention, along with a fine of 2,000 RMB.
Model Output	The error type is Prison Term Error.
Ground Truth	The error type is Charge Error.
Error Analysis	The defendant violently assaulted a police officer who was lawfully performing his duties, thereby committing the crime of obstructing official duties; by driving a motor vehicle on a public road while intoxicated, his actions constituted the crime of dangerous driving. The model failed to identify the root error (the charge error) in the chain of errors, instead classifying it as a prison term error.

Table 16: A failure case in the misattribution of root causes in cascading errors.

Key Fact Snippet	... In 2009, Yang XX borrowed 70,000 RMB from the victim Shi XX through the defendant Hou XX using a resettlement housing contract; subsequently, Yang XX borrowed an additional 40,000 RMB from Shi XX in installments...
Erroneous Judgment	The defendant Hou XX was convicted of fraud and sentenced to three years and nine months in prison, along with a fine of 80,000 RMB.
Model Output	The defendant should be convicted of the crime of fraud.
Ground Truth	The defendant should be convicted of the crime of contract fraud.
Error Analysis	Even when the prompt explicitly indicates that there is a charge error in the judgment, the model still repeats the incorrect charge from the erroneous judgment.

Table 17: A failure case showing anchoring bias.

Key Fact Snippet	Since May 2013, the defendant Qiu XX has been selling various types of illegally manufactured invoices for the purpose of obtaining illegal profits. On August 1, police officers... apprehended Qiu and seized 6,536 illegally manufactured invoices of various types (which, upon examination, were found to include 3,284 local tax invoices and 108 special VAT invoice forms, all of which were counterfeit), as well as a batch of invoice stubs and express delivery waybill stubs, along with counterfeiting tools such as computers and printers.
Erroneous Judgment	The defendant, Qiu XX, was convicted of the crime of illegally manufacturing or selling illegally manufactured invoices and was sentenced to two years in prison and fined 9,077 RMB.
Model Output	The judgment contains a charge error.
Ground Truth	The judgment contains a fine error.
Error Analysis	Since the crime of illegally manufacturing or selling illegally manufactured invoices is a rare offense, we surmise that it appears infrequently in the model's training corpus; therefore, the model classified it as a charge error rather than a fine error.

Table 18: A failure case in the diagnostic logic misalignment in long-tail charges.

Models	Error Detection	Error Classification	Error Correction
GPT-5	100.00	98.13	99.98
Qwen3-Max	100.00	100.00	100.00
DeepSeek-V3.2-Chat	99.95	99.92	99.97
GLM-4.6-Chat	99.97	87.57	98.81
MiniMax-M2	100.00	100.00	100.00
Qwen3-14B-Chat	100.00	100.00	100.00
Qwen3-8B-Chat	100.00	99.84	99.57
GPT-4o	100.00	99.79	100.00
Qwen3-Max-preview-Thinking	99.70	99.70	99.97
DeepSeek-V3.2-Thinking	99.85	96.44	99.65
Kimi-K2-Thinking	99.34	98.61	99.73
Qwen3-Next-80B-A3B-Thinking	99.94	99.91	99.94
Farui-Plus	99.85	99.76	99.91
LawLLM-7B	100.00	100.00	100.00

Table 19: Instruction Following Rate (%) on three subtasks of AR-BENCH.

Prompt for Error Detection.

Role and Mission

You are a professional criminal legal analysis expert specializing in determining whether a judgment contains errors based on case facts (Fact), the reasoning process and legal provisions cited in the original judgment (Reasoning), and the judgment itself (Judgment).

Core Task

Based on the user-provided “Case Facts (Fact)”, “Original Judgment’s Reasoning Process and Legal Basis (Reasoning)”, and “Original Judgment (Judgment)”, accomplish one thing:

1. Determine whether the original judgment contains errors (output only true/false);

Analysis Basis

1. Strictly adhere to relevant provisions of China’s Criminal Law, integrating core elements from the case facts (e.g., criminal acts, amounts involved, circumstances of the crime, restitution status, etc.);
2. Focus on examining whether the reasoning process and legal provisions cited in the original judgment (Reasoning) are logically coherent, legally accurate, and form a complete closed loop with the case facts;
3. Error criteria: An explicit conflict between the original judgment/reasoning/cited legal provisions and case facts/legal statutes, excluding reasonable sentencing range variations;
4. Focus solely on the three core elements—“Charge, Prison Term, and Fine”—excluding other procedural content in the judgment (e.g., appeal deadlines, restitution methods).

Output Format (Strictly required; otherwise deemed invalid) Output only a JSON string containing one field: llm_iserror (bool). Example:

```
{
  "llm_iserror": true
}
or
{
  "llm_iserror": false
}
```