ChJGBench: A Comprehensive Benchmark for Chinese Legal Judgment Generation with Diagnostic Evaluation

Anonymous ACL submission

Abstract

With the rapid development of LLMs, increasing attention has been paid to whether they can reliably generate complete and legally valid judicial documents. This calls for a benchmark capable of evaluating whether models can accurately apply statutes and resolve outcome-critical issues, including claims, charges, penalties, and monetary obligations. However, existing benchmarks typically focus on isolated legal tasks or coarse-grained evaluation. We introduce ChJGBench, a benchmark for Chinese legal judgment generation covering civil, criminal, and administrative cases. Built from 3,462 real-world judgments, ChJGBench provides a case-type-aware diagnostic framework comprising 13 substantive dimensions, together with a hybrid rule- and LLM-based evaluation paradigm. Experiments show that specialized legal models do not consistently outperform large general-purpose models, and their advantages vary across tasks. Criminal cases achieve the best overall performance, whereas administrative cases remain the most challenging. ***The related resources will be released upon acceptance of the paper.***

1 Introduction

In China, written judgments are the primary official records through which courts present factual findings, legal reasoning, and dispositive outcomes. Drafting such judgments requires models to not only generate fluent legal text, but also correctly resolve outcome-critical legal elements, such as claims, charges, penalties, monetary obligations, and statutory grounds. Although large language models (LLMs) have shown promise in legal text generation, it remains unclear whether they can reliably generate complete and legally valid judgments across diverse litigation scenarios.

Existing legal benchmarks (Fei et al., 2023; Chalkidis et al., 2022; Guha et al., 2023; Li et al., 2024) have contributed to the systematic evaluation of legal understanding and reasoning, but they remain insufficient for evaluating judicial judgment generation. As summarized in Table 1, prior benchmarks are often limited in at least one of the following aspects: litigation-type coverage, evaluation scale, diagnostic granularity, or evaluation methodology. Many benchmarks focus on isolated legal tasks, such as classification, multiple-choice question answering, charge prediction, or legal consultation. Even recent benchmarks on judgment or legal document generation typically cover fewer samples, fewer evaluation dimensions, or only partial litigation types. Moreover, commonly used text-similarity metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are inadequate for legal evaluation, as surface-level overlap may overlook substantive legal errors, including incorrect charges, missing payment obligations, or wrong statutory provisions. Unconstrained LLM-as-a-judge evaluation may further introduce subjectivity and instability when assessing long and complex judicial documents.

Benchmark	M-Type	#Samp.	#Dim.	Eval. Method
JuDGE	×	501	4	Det.
CaseGen	✓	500	4	LLM
J1-Bench	×	162	6	Det. + LLM
PLawBench	?	40	6	Rubric + LLM
ChJGBench	✓	3,462	13	Diag. + LLM

Table 1: Comparison between **ChJGBench** and existing legal judgment/document generation benchmarks. **M-Type**: coverage of criminal, civil, and administrative cases; **#Samp.**: evaluation samples for judgment/document generation; **#Dim.**: substantive evaluation dimensions; **Det.**: deterministic evaluation, e.g., exact, pattern, or regex matching; **Diag.**: diagnosis-aware evaluation with litigation-type-specific criteria.

To address these limitations, we introduce **ChJGBench**, a comprehensive benchmark for

Abbr.	Dimension	Section
Law	Statutory Provision Application	4.3
ISS	Disputed Issues Identification	4.3
JUS	Judicial Justification	4.3
Charge	Criminal Charge Prediction	4.4
Fine_R	Monetary Fine Recall	4.4
REC	Recidivism Assessment	4.4
MIT	Mitigating Circumstances	4.4
Pen_M	Penalty Duration Prediction	4.4
SUP$_{\mathrm{Civ}}$	Claim Support	4.5
PAY_R$_{\mathrm{Civ}}$	Payment Obligation Recall	4.5
SUP$_{\mathrm{Adm}}$	Claim Support	4.6
PAY_R$_{\mathrm{Adm}}$	Payment Obligation Recall	4.6
DIS	Administrative Disposition	4.6

Table 2: Case-type-aware evaluation dimensions in ChJGBench, grouped by background color into common, criminal, civil, and administrative dimensions. Each entry reports the metric abbreviation and points to the section where the metric is defined.

Chinese legal judgment generation. ChJGBench contains 3,462 real-world judgments issued between 2019 and 2024, including 1,650 civil cases, 1,022 criminal cases, and 790 administrative cases. Rather than directly sampling from the empirical distribution of collected judgments, which would overrepresent frequent categories and underrepresent long-tail case types, we adopt a coverage-oriented sampling strategy that balances distributional fidelity with category coverage. This design allows ChJGBench to better reflect both common and less frequent adjudicative scenarios.

A key feature of ChJGBench is its case-type-aware diagnostic evaluation framework. Instead of assigning a single generic score to generated judgments, we evaluate them along both common and litigation-type-specific dimensions. The common dimensions assess statutory provision application, disputed issue identification, and judicial justification. The case-specific dimensions capture distinctive adjudicative requirements in different litigation types: criminal cases involve charge prediction, monetary fine recall, recidivism assessment, mitigating circumstance identification, and penalty duration prediction; civil cases involve claim support and payment obligation recall; and administrative cases involve claim support, administrative disposition, and payment obligation recall. In total, ChJGBench provides 13 substantive evaluation dimensions, enabling fine-grained diagnosis of both shared legal reasoning capabilities and litigation-specific adjudicative outcomes.

To improve evaluation reliability and scalability, ChJGBench adopts a hybrid metric computation paradigm. LLMs are used in a constrained manner to extract or normalize legal elements, such as cited statutory provisions, claim-support outcomes, monetary amounts, and punishment terms. These elements are then compared with reference judgments through deterministic matching, numerical deviation, or semantic matching. In this way, LLMs serve as controlled information extractors rather than unconstrained subjective judges, combining the objectivity of rule-based evaluation with the flexibility needed for legal semantic comparison.

Our contributions are threefold. First, we construct **ChJGBench**, a large-scale benchmark for Chinese legal judgment generation with 3,462 real-world judgments and a coverage-oriented data distribution across civil, criminal, and administrative cases. Second, we propose a case-type-aware diagnostic evaluation framework with 13 fine-grained dimensions tailored to common legal reasoning and litigation-specific adjudicative requirements. Third, we conduct comprehensive experiments on ChJGBench to analyze model performance across litigation types, legal task families, and diagnostic dimensions, providing empirical insights for future research on reliable legal judgment generation.

2 Related Work

2.1 Legal Benchmarks

With the rapid adoption of LLMs in the legal domain (Kang et al., 2019; Louis et al., 2023), there is an increasing demand for benchmarks that better reflect real-world legal practice. Existing efforts have primarily focused on evaluating legal understanding and reasoning through structured tasks. For example, LexGLUE (Chalkidis et al., 2022) provides a suite of multi-label classification and multiple-choice tasks derived from European and U.S. legal systems. LegalBench (Guha et al., 2023) and LEXam (Fan et al., 2026) further extend evaluation within case law systems, covering diverse forms of legal reasoning such as rule application and legal interpretation. In the Chinese legal domain, benchmarks such as LawBench (Fei et al., 2023) and LexEval (Li et al., 2024) adopt a multi-task paradigm to assess a broad spectrum of capabilities, including charge prediction, simulated bar exam questions, and legal consultation.

These benchmarks primarily focus on natural language understanding, classification, or question answering tasks within the legal domain. However, real-world judicial practice requires more than solving isolated legal tasks: courts must produce complete decisions that integrate fact finding, legal reasoning, and final dispositions. This mismatch highlights the need for benchmarks that evaluate legal judgment generation in a comprehensive and structured manner.

2.2 Legal Judgment Document Generation

Several recent studies have examined how to evaluate the generation of complete legal judgments. For instance, the CAIL 2024 judgment generation task used ROUGE-L (Lin, 2004) to assess factual accuracy through extracted events and BERTScore (Zhang et al., 2020) to measure semantic consistency. However, such metrics capture only coarse textual similarity and often fail to distinguish similar charges or key sentencing outcomes. Recent benchmarks have moved toward more structured and legally informed evaluation. JuDGE (Su et al., 2025) evaluates Chinese criminal judgment generation with dimensions such as charge prediction, penalty accuracy, and legal reference consistency, but is limited to criminal cases. CaseGen (Li et al., 2025) builds an expert-annotated dataset from 500 real cases and adopts LLM-as-a-judge for evaluation. J1-Bench (Jia et al., 2026) and PLawBench (Shi et al., 2026) further assess legal drafting in broader practical legal scenarios, combining rule-based checks, expert-designed rubrics, and LLM-based scoring.

Despite these efforts, existing approaches face limitations: similarity-based metrics overlook legal correctness, while rubric-based LLM evaluation may introduce subjectivity and scalability concerns. In addition, most benchmarks either focus on a single case type or treat judgment drafting as one subtask within legal capability evaluation. In contrast, our work targets Chinese judgment generation across criminal, civil, and administrative cases, and combines exact matching with LLM-based semantic evaluation to assess outcome correctness and judgment quality in a unified framework.

3 Task Definition

Chinese judicial documents include judgments, rulings, mediation statements, decisions, notices, orders, and other instruments. Among them, **judgments** constitute the most representative and practically important category of judicial decisions, as they require enforceable rulings, rigorous legal reasoning, and a standardized document structure. We therefore focus on judgment generation as the core task. To enable systematic modeling and evaluation, we decompose a Chinese court judgment into six logical components and formally define the task accordingly. Figure 1 illustrates this structure, and further details are provided in Appendix A.

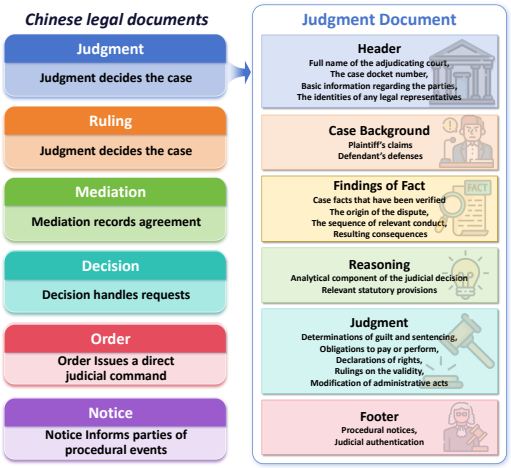

Figure 1: Structure of a legal judgment document.

$$\mathcal{D} = (Header,\ Background,\ Findings\ of\ Fact,\ Reasoning,\ Judgment,\ Footer) \quad (1)$$

The first three components provide the factual and procedural basis of the case and are largely extractive or descriptive in nature, making their generation manageable for contemporary LLMs compared with the reasoning-intensive components. The *Footer*, being largely template-based, is likewise excluded from the primary generation targets.

By contrast, the *Reasoning* and *Judgment* sections constitute the core of a judicial decision. The former reflects the legal reasoning process, namely how legal conclusions are derived from established facts, while the latter presents the final adjudicative ruling. Generating these sections requires the model to integrate legal knowledge, perform legal inference, and conform to established conventions of judicial practice, thereby enabling a reliable mapping from factual descriptions to adjudicative outcomes. Thus, these two sections serve as the primary generation targets in our task.

In summary, the legal judgment generation task

can be formally defined as follows: given

$$\mathcal{X} = (Header,\ Background,\ Findings\ of\ Fact),$$

the model is required to generate

$$\mathcal{Y} = (Reasoning,\ Judgment).$$

4 The ChJGBench Benchmark

4.1 Benchmark Construction

Data Sources We construct ChJGBench to evaluate Chinese legal judgment generation across civil, criminal, and administrative cases. All judgments are collected from China Judgments Online[1], the official national repository of Chinese court documents. Each judgment is parsed into six standard structural components (see Figure 1), supporting flexible input construction.

Data Distribution Sampling directly from the empirical distribution of collected judgments would overrepresent frequent categories and underrepresent long-tail ones. To balance distributional fidelity with category coverage, we adopt a coverage-oriented sampling strategy. We first stratify the candidate pool by case type t and issuance year y from 2019 to 2024. Within each stratum, cases are sampled according to legal category c: cause of action for civil cases, charge for criminal cases, and dispute category for administrative cases.

Let $P_{t,y}(c)$ denote the empirical category distribution within stratum (t, y). To improve long-tail coverage, we apply a power-temperature transformation:

$$\tilde{P}_\tau(c \mid t, y) = \frac{P_{t,y}(c)^\tau}{\sum_k P_{t,y}(k)^\tau}, \quad \tau \in (0, 1]. \quad (2)$$

When $\tau < 1$, low-frequency categories receive higher relative sampling probabilities, while dominant categories are down-weighted.

To constrain deviation from the empirical distribution, we interpolate the original and smoothed distributions:

$$Q_{\lambda,\tau}(c \mid t, y) = \lambda P_{t,y}(c) + (1 - \lambda)\tilde{P}_\tau(c \mid t, y), \quad (3)$$

where $\lambda \in [0, 1]$ balances empirical fidelity and long-tail coverage. Instances are sampled according to $Q_{\lambda,\tau}$ under the target size of each case type.

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

After structural parsing, we apply quality filters to retain only formal judgment documents, removing duplicates, documents with incomplete structural fields, and cases with failed or ambiguous section segmentation. To ensure sufficient factual and legal content, we require both the *Findings of Fact* and *Court's Reasoning* sections to contain more than 100 Chinese characters. The final benchmark contains 3,462 judgments issued between 2019 and 2024, including 1,650 civil, 1,022 criminal, and 790 administrative cases. Detailed dataset statistics are provided in Appendix D.

4.2 Case-Type-Aware Diagnostic Evaluation Framework

A central challenge in evaluating legal judgment generation is that a judgment is not simply a fluent piece of legal writing. It is a structured legal decision that must identify disputed issues, apply relevant statutory provisions, provide judicial reasoning, and reach legally valid outcomes. Existing evaluation methods are often insufficient for this setting. Surface-level similarity metrics, such as lexical overlap or embedding similarity, may fail to detect legally decisive errors; for example, a judgment with a correct format but incorrect charges or statutory provisions may still receive a high score. Meanwhile, directly asking an LLM to assign a quality score, even with clear instructions, inevitably introduces subjectivity and instability, especially for long and complex judgments.

The key design of our framework is to combine LLM-based extraction with objective evaluation metrics. For each generated judgment, LLMs are used only to extract or normalize specific legal elements under explicit prompts, such as whether a claim is supported, which statutory provisions are cited, or what sentence length is imposed. The extracted elements are then evaluated using deterministic matching, numerical deviation, or semantic matching against the reference judgment. In this way, LLMs do not serve as unconstrained subjective judges; instead, they function as controlled information extractors that enable scalable and fine-grained legal evaluation.

This design makes the evaluation both diagnostic and legally grounded. It reveals not only whether a generated judgment resembles the reference text, but also whether it reaches correct conclusions on key adjudicative components. Table 2 summarizes the resulting evaluation dimensions.

Instead of applying a single uniform metric set

to all judgments, we decompose legal judgment generation into two levels of evaluation dimensions. The first level consists of dimensions that are common to civil, criminal, and administrative cases, including statutory provision application, legal justification, and disputed issue identification. The second level consists of case-type-specific dimensions that reflect the distinctive adjudicative tasks of criminal, civil, and administrative litigation.

4.3 Shared Dimensions

Statutory Provision Application (Law). This metric evaluates whether the statutory provisions cited in the model-generated "Court's Reasoning" section match those in the reference judgment, and is applicable to civil, criminal, and administrative cases. We normalize both the predicted and reference provisions to the format `[Statute Name] Article X`, and then compute F_1 score by exact matching the two provision lists.

Disputed Issues Identification (ISS). This metric evaluates whether the generated judgment captures the core legal questions under adjudication. These issues vary by case type: they may involve rights and obligations between equal parties in civil cases, criminal prosecution by the state in criminal cases, and the legality of administrative actions in administrative cases. We extract disputed issues from both the reference and generated judgments using an LLM. An evaluator LLM then compares the two lists under a predefined rubric and assigns an issue-alignment score, indicating how well the generated judgment covers the legally salient disputes in the reference judgment.

Judicial Justification (JUS). This metric evaluates whether the generated judgment provides a logically coherent, legally sound, and detailed rationale that connects the facts, applicable law, disputed issues, and final disposition. Unlike general text-generation evaluation, this metric focuses on the adequacy of legal reasoning rather than linguistic fluency. We use an LLM to extract the key reasoning points from both the reference judgment and the generated judgment. An evaluator LLM then compares the two extracted lists under a predefined rubric and assigns a justification score.

4.4 Criminal Case Dimensions

Criminal Charge Prediction (Charge). This metric evaluates criminal charge prediction by comparing the charges extracted from the generated and reference judgments. We use an LLM to extract charge labels and compute the F_1 score through label matching.

Monetary Fine Recall (Fine_R). This metric evaluates whether the generated judgment accurately recalls monetary fines imposed in criminal cases. Fines are an important component of criminal sanctions, especially in cases involving economic crimes, property crimes, or offenses for which financial punishment is statutorily prescribed. To evaluate fine amounts, we use the Normalized Absolute Deviation (NAD). For each monetary fine, let y_{pred} denote the predicted amount and y_{true} denote the reference amount. NAD is defined as:

$$\text{NAD} = \begin{cases} 1, & \text{if } y_{\text{pred}} = 0 \text{ and } y_{\text{true}} = 0, \\ 1 - \dfrac{|y_{\text{pred}} - y_{\text{true}}|}{\max(y_{\text{pred}}, y_{\text{true}})}, & \text{otherwise.} \end{cases} \tag{4}$$

Evaluation is restricted to criminal cases where the reference judgment contains a non-zero fine.

Recidivism Assessment (REC). This metric evaluates whether the generated judgment correctly identifies the defendant's recidivist status. Recidivism is a legally significant sentencing factor because it may affect sentencing severity and the availability of lenient treatment. We formulate this dimension as a binary classification task, compare the predicted label with the reference label, and report F_1 score.

Mitigating-Circumstance Identification (MIT). This metric evaluates whether the generated judgment correctly identifies sentencing factors that justify leniency. Such circumstances may include voluntary surrender, confession, remorse, restitution, compensation, or obtaining the victim's forgiveness. We treat this dimension as a structured extraction and binary evaluation task, and measure performance using F_1 score.

Penalty Duration Prediction (Pen_M). This metric evaluates whether the generated judgment predicts the correct duration of the principal punishment. It applies to criminal cases involving duration-based penalties, such as fixed-term imprisonment, criminal detention, or public surveillance. For cases with a reference penalty duration, we compute the NAD score defined in Equation 4, where y_{pred} and y_{true} denote the predicted and reference penalty durations, respectively. The final

penalty-duration score (PDS) is the average NAD across all applicable criminal cases:

$$\mathrm{PDS} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{NAD}(y_{\text{pred}}^{(i)}, y_{\text{true}}^{(i)}), \tag{5}$$

where N is the number of criminal cases with duration-based penalties.

4.5 Civil Case Dimensions

Claim Support ($\mathrm{SUP}_{\mathrm{Civ}}$). This metric evaluates whether the generated civil judgment correctly determines whether the plaintiff's claims are upheld, partially upheld, or rejected. It captures the substantive adjudicative outcome of civil litigation. We measure performance using F_1 score based on the alignment between the generated and reference claim-support structures.

Payment Obligation Recall ($\mathrm{PAY_R}_{\mathrm{Civ}}$). This metric evaluates whether the generated civil judgment accurately recalls payment obligations, such as compensation, repayment, damages, restitution, or other monetary liabilities. We extract payment amounts from both the generated and reference judgments and evaluate numerical accuracy using NAD, as defined in Equation 4. Evaluation is restricted to civil cases where the reference judgment contains a non-zero payment obligation.

4.6 Administrative Case Dimensions

Claim Support ($\mathrm{SUP}_{\mathrm{Adm}}$). This metric evaluates whether the generated administrative judgment correctly determines the outcome of the plaintiff's claims. Although this dimension is also used in civil cases, its legal meaning differs in administrative litigation, where claims typically concern the legality of administrative acts, administrative nonfeasance, or requests for administrative remedies. We extract supported and rejected claims from both the generated and reference judgments, and compute F_1 score to quantify the alignment between the two claim-support structures.

Payment Obligation Recall ($\mathrm{PAY_R}_{\mathrm{Adm}}$). This metric evaluates whether the generated administrative judgment accurately recalls monetary obligations, such as administrative compensation, restitution, refund obligations, or other payment-related remedies. As in civil cases, we evaluate numerical accuracy using NAD. Evaluation is restricted to administrative cases where the reference judgment contains a non-zero payment obligation.

Administrative Disposition Prediction (DIS). This metric evaluates whether the generated administrative judgment accurately predicts non-monetary administrative dispositions, such as annulment of an administrative act, orders to reissue a decision, or declarations of unlawfulness. We extract the core disposition from both the generated and reference judgments, such as "revoked", "upheld", or "remanded", and evaluate their semantic alignment. Evaluation is restricted to administrative cases with explicit dispositive language in the reference judgment.

5 Experiments

This section presents a comprehensive evaluation of current LLMs on **ChJGBench**. We compare legal-domain models and strong general-purpose LLMs across civil, criminal, and administrative cases under the same diagnostic evaluation framework. Table 3 summarizes the main results over all key metrics. We then introduce the experimental settings, discuss major performance trends, and analyze task difficulty and cross-model variability across evaluation dimensions.

5.1 Experimental Settings

Models. We evaluate both legal-domain models and strong general-purpose LLMs on judicial-document completion tasks. All models are evaluated under a few-shot prompting setting. The evaluated models include two legal-domain systems, wisdomInterrogatory (Wu et al., 2026) and LegalOne (Li et al., 2026), as well as seven general-purpose models: Claude Sonnet 4.5 (Anthropic, 2025), DeepSeek-R1 (Guo et al., 2025), Gemini-3 Pro Preview (Google DeepMind, 2026), GPT-5.4 (OpenAI, 2026), Grok-4 (xAI, 2025), LLaMA-3.1 405B (Grattafiori et al., 2024), and Qwen-max (Yang et al., 2025).

Metrics. The evaluation dimensions are summarized in Table 2. For classification-based tasks, we adopt the F_1 score; for numerical estimation tasks, we adopt NAD-Recall. All reported scores are rounded to three decimal places.

Evaluation Model Validation. To assess model outputs in a consistent manner, we use Qwen-Plus as the evaluation model for information extraction and necessary semantic comparison. The extraction accuracy of Qwen-Plus and its agreement with human annotations in semantic comparison are

Case / Metric	Model								
	WisdomI.	**LegalOne**	**Claude Sonnet 4.5**	**DS-R1**	**Gemini 3-Pro**	**GPT 5.4**	**Grok 4**	**LLaMA 405B**	**Qwen max**
Criminal									
Law	0.388	0.711	0.678	0.684	**0.747**	0.645	0.610	0.510	0.645
JUS	0.387	0.714	0.738	**0.760**	0.737	0.703	0.740	0.643	0.734
ISS	0.513	0.799	0.810	0.817	0.823	0.786	**0.826**	0.707	0.799
Charge	0.436	0.973	0.970	0.969	0.971	0.975	**0.976**	0.957	0.968
Fine_R	0.135	0.565	0.577	0.563	**0.611**	0.573	0.501	0.476	0.595
REC	0.108	0.743	0.705	0.715	**0.753**	0.718	0.668	0.500	0.715
MIT	0.408	0.967	0.968	0.961	0.973	0.963	**0.973**	0.885	0.972
Pen_M	0.608	0.731	0.751	0.740	**0.786**	0.748	0.724	0.602	0.761
Avg	0.373	0.775	0.775	0.776	**0.800**	0.764	0.752	0.660	0.774
Civil									
Law	0.080	0.422	0.441	0.231	**0.593**	0.465	0.395	0.172	0.218
JUS	0.422	0.584	**0.720**	0.697	0.654	0.632	0.713	0.516	0.651
ISS	0.633	0.740	0.816	0.823	0.810	0.759	0.818	0.743	**0.828**
SUP_{Civ}	0.649	0.786	0.779	0.783	**0.831**	0.802	0.761	0.682	0.742
$\text{PAY_R}_{\text{Civ}}$	0.244	0.742	0.743	0.753	**0.757**	0.748	0.709	0.616	0.736
Avg	0.406	0.655	0.700	0.657	**0.729**	0.681	0.679	0.546	0.635
Administrative									
Law	0.107	0.376	0.564	0.542	**0.579**	0.514	0.546	0.257	0.429
JUS	0.309	0.589	**0.737**	0.717	0.668	0.634	0.735	0.524	0.688
ISS	0.536	0.722	0.798	**0.805**	0.790	0.745	0.798	0.689	0.802
SUP_{Adm}	0.686	0.810	0.823	0.844	**0.851**	0.827	0.825	0.744	0.830
$\text{PAY_R}_{\text{Adm}}$	0.035	0.353	0.369	0.218	**0.440**	0.432	0.305	0.226	0.314
DIS	0.326	0.577	0.764	0.787	0.731	0.698	**0.799**	0.712	0.682
Avg	0.333	0.571	0.676	0.652	**0.677**	0.642	0.668	0.525	0.624

Table 3: Comparison of key metrics across models by case type. Unless otherwise specified, we report F1 scores for classification tasks and NAD-Recall for amount-related tasks. Red bold cells indicate the best value in each row, and blue underlined cells indicate the second-best value; ties are assigned the same rank. Abbreviations: WisdomI. = wisdomInterrogatory; DS-R1 = DeepSeek-R1.

reported in Appendix B.1 and Appendix B.2, respectively. Qwen-Plus achieves an average extraction accuracy of 99.41% across three independent runs, demonstrating reliable and stable structured extraction performance. For semantic matching, Qwen-Plus shows strong agreement with human annotations, with a Cohen's κ of 0.7727 and a Krippendorff's α of 0.7780. These validation results support the use of Qwen-Plus as a reliable evaluation tool for both rule-based information extraction and legal semantic comparison in ChJGBench.

Experimental Details. We use one-shot prompting for all tasks. To ensure fairness and comparability, we keep the prompt template, demonstration example, input context, and output constraints identical across all models. All inference is performed with a decoding temperature of 0.

5.2 Main Results

Specialized legal models do not necessarily outperform large-scale general models. LegalOne, a recent 8B legal-domain model, performs competitively on criminal tasks. However, in civil and administrative cases, it is often surpassed by general models such as Gemini 3-Pro-Preview and Claude Sonnet 4.5. This suggests that training small-scale models on legal-domain data can inject legal knowledge and improve performance on certain tasks, but such models still fall short of mainstream large-scale models on complex legal reasoning and numerical judgment.

Model advantages are task-specific rather than universal. Although Gemini 3-Pro-Preview obtains the highest average scores across criminal, civil, and administrative cases, there remains room for improvement across many evaluation dimensions. DeepSeek-R1 leads on criminal JUS and administrative ISS, while Grok-4 shows strong performance on criminal ISS, Charge, and MIT, as well as administrative DIS. Claude Sonnet 4.5 performs best on civil and administrative JUS, and Qwen-max obtains the highest score on civil ISS. This pattern suggests that different models capture different aspects of legal judgment completion. Some models show stronger capabilities in JUS and ISS, whereas others are more effective in predicting

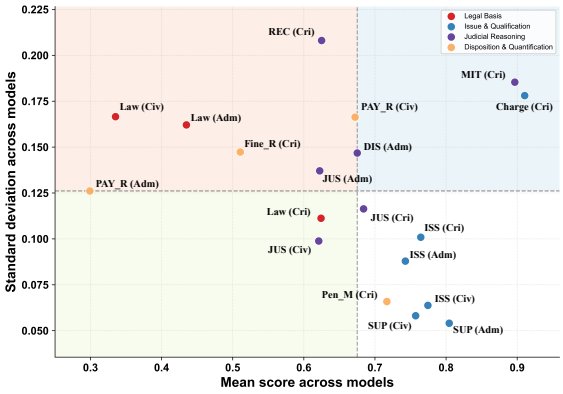

Figure 2: Task difficulty and cross-model disagreement across evaluation dimensions. Each point denotes a case-type-specific task. The dashed vertical and horizontal lines denote the median mean score and the median standard deviation, respectively. Point colors indicate the legal ability family to which each task belongs; the family taxonomy is provided in Appendix C.

concrete legal outcomes.

Criminal cases achieve the highest overall performance, while administrative cases are the most challenging. The results show a clear performance gap across case types: criminal cases achieve the highest average scores, civil cases rank in the middle, and administrative cases are the most challenging. One plausible explanation is that criminal cases involve more stable mappings among Charge, sentencing factors, MIT, and principal penalties, as reflected by the high scores on Charge and MIT and the relatively strong performance on Pen_M. By contrast, civil and administrative cases require more open-ended fact finding, rights-and-obligations assessment, and statutory application.

5.3 Analysis of Task Difficulty and Cross-Model Variability

The four quadrants reflect different evaluation implications. The left-side quadrants indicate common limitations of current models, as tasks in these regions have relatively low mean scores. From a task-family perspective, ***Legal Basis*** is the most evident weakness: all Law-related dimensions are located on the left side, showing that statutory grounding remains difficult across case types. This is especially pronounced for Law (Civ) and Law (Adm), suggesting that models struggle to accurately ground generated judgments in domain-specific legal norms. In addition, PAY_R (Adm), Fine_R (Cri), and REC (Cri) also appear on the left side, indicating that payment obligation recall, fine-related reasoning, and recidivism assessment remain challenging adjudicative tasks.

The upper-right quadrant contains tasks with relatively high mean scores but large cross-model variance. These tasks are not uniformly difficult, as some models can already perform well on them; however, they remain **highly discriminative**. In other words, they serve as high-quality benchmark indicators. Specifically, metrics such as Charge (Cri) and MIT (Cri) can effectively distinguish models with stronger legal qualification and adjudicative reasoning capabilities.

In contrast, the lower-right quadrant includes high-mean, low-variance dimensions, such as ISS across all case types and SUP in civil and administrative cases. These tasks appear comparatively easier for current models, suggesting that models can generally identify disputed issues and claim support. However, their low variance makes them **less effective at differentiating the capabilities** of advanced models.

6 Conclusions

In this paper, we introduced **ChJGBench**, a comprehensive and diagnostically grounded benchmark for Chinese legal judgment generation. Built from 3,462 diverse real-world judgments across civil, criminal, and administrative cases, ChJGBench adopts a coverage-oriented sampling strategy to capture both common and long-tail adjudicative scenarios. It further provides a case-type-aware diagnostic evaluation framework with 13 substantive dimensions and a hybrid rule-and-LLM evaluation paradigm, enabling fine-grained, interpretable, and reliable assessment of both shared legal reasoning abilities and litigation-specific adjudicative outcomes. Our experiments provide several key insights into the current capabilities of LLMs for Chinese legal judgment generation. First, specialized legal models do not consistently outperform large-scale general-purpose models, indicating that legal-domain training alone is not sufficient for complex judgment generation. Second, model strengths are task-specific rather than universal, with different models excelling on different diagnostic metrics. Third, performance varies substantially across litigation types: criminal cases achieve the highest overall scores, whereas administrative cases remain the most challenging due to the complexity of fact finding, rights-and-obligations assessment, and statutory application.

Limitations

ChJGBench has several limitations. First, it is limited to mainland China's legal system and may not transfer to other jurisdictions with different legal doctrines, procedures, or judgment-writing conventions. Second, since the benchmark is constructed from publicly available court judgments, it may not fully represent the distribution of real-world legal disputes, particularly cases that are not publicly disclosed. This may introduce selection bias into the dataset. Third, the benchmark is static, whereas laws, judicial interpretations, and adjudicatory standards may change over time. Therefore, performance on ChJGBench should be understood as a research-oriented measure of model capability, rather than as evidence that language models are suitable for real-world judicial decision-making.

Ethical Statement

ChJGBench is constructed from publicly available Chinese court judgments, with personally identifiable information removed to protect privacy and reduce re-identification risks. The benchmark is designed solely for research purposes, aiming to evaluate the technical capabilities and limitations of language models in legal judgment generation. The dataset must not be used to support or justify the deployment of AI systems in real-world judicial decision-making without rigorous human oversight, legal validation, institutional accountability, and compliance with applicable laws and regulations. We acknowledge that judicial data may reflect historical or social biases, which models may reproduce.

We used AI tools only for language polishing and grammar refinement. AI tools were not used to design the experiments, develop the methodology, construct the benchmark, analyze the results, or draw conclusions. All scientific contributions, experimental design, implementation, analysis, and writing decisions were made by the authors.

References

Anthropic. 2025. Claude Sonnet 4.5 System Card.

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

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. *Preprint*, arXiv:2505.12864.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *Preprint*, arXiv:2309.16289.

Google DeepMind. 2026. Gemini 3 Pro Model Card. Model release: November 2025; last updated: May 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Zheng Jia, Shengbin Yue, Wei Chen, Siyuan Wang, Yidong Liu, Zejun Li, Yun Song, and Zhongyu Wei. 2026. Ready jurist one: Benchmarking language agents for legal intelligence in dynamic environments. *Preprint*, arXiv:2507.04037.

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

Haitao Li, Yifan Chen, Shuo Miao, Qian Dong, Jia Chen, Yiran Hu, Junjie Chen, Minghao Qin, Yueyue Wu, Yujia Zhou, Qingyao Ai, Yiqun Liu, Cheng Luo, Quan Zhou, Ya Zhang, and Jikun Hu. 2026. Legalone: A family of foundation models for reliable legal reasoning. *Preprint*, arXiv:2602.00642.

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. *Preprint*, arXiv:2409.20288.

Haitao Li, Jiaying Ye, Yiran Hu, Jia Chen, Qingyao Ai, Yueyue Wu, Junjie Chen, Yifan Chen, Cheng Luo, Quan Zhou, and Yiqun Liu. 2025. Casegen: A benchmark for multi-stage legal case documents generation. *Preprint*, arXiv:2502.17943.

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

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023. Interpretable long-form legal question answering with retrieval-augmented large language models. *Preprint*, arXiv:2309.17050.

OpenAI. 2026. Introducing GPT-5.4.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yuzhen Shi, Huanghai Liu, Yiran Hu, Gaojie Song, Xinran Xu, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Di Feng, Wenbo Lv, Weiheng Wu, Kexin Yang, Sen Yang, Wei Wang, Rongyao Shi, Yuanyang Qiu, Yuemeng Qi, Jingwen Zhang, and 11 others. 2026. Plawbench: A rubric-based benchmark for evaluating llms in real-world legal practice. *Preprint*, arXiv:2601.16669.

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. *Preprint*, arXiv:2503.14258.

Yiquan Wu, Yuhang Liu, Yifei Liu, Ang Li, Siying Zhou, Kun Kuang, and Fei Wu. 2026. Wisdominterrogatory (luwen): An open-source legal large language model technical report. *Preprint*, arXiv:2604.06737.

xAI. 2025. Grok 4 Model Card. Last updated: August 20, 2025.

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. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Judgment details

- *Header*: Contains basic procedural information, including the court name, case number, and the identities of the parties and their representatives.
- *Background*: Summarizes the claims, defenses, or charges raised by the parties at the outset of the litigation.
- *Findings of Fact*: Presents the facts established through evidence, along with a brief assessment of the supporting evidence.
- *Reasoning*: Provides the court's legal analysis by applying relevant statutes to the facts and evaluating the parties' arguments.
- *Judgment*: States the court's final and binding ruling in a clear and enforceable format.
- *Footer*: Includes procedural notices, such as appeal rights, judicial signatures, the date, and the court seal.

Header constitutes the procedural opening of a judicial decision, primarily setting forth the full name of the adjudicating court, the case docket number, basic information regarding the parties, and the identities of any legal representatives. This section serves to establish the jurisdictional basis, the level of adjudication, and the lawful standing of all participants in the proceedings. It reflects the formal entry of the dispute into the judicial process without engaging with any substantive legal issues. As a purely formal component, it provides the procedural framework necessary for subsequent factual findings and legal analysis.

Case Background consolidates the core positions advanced by the parties during litigation, specifically the plaintiff's claims and the defendant's defenses. In criminal or administrative cases, this corresponds respectively to the prosecutorial charges or the administrative authority's enforcement rationale on one side, and the defendant's or administrative counterpart's explanations and objections on the other. This part delineates the contours of the dispute, thereby defining the scope within which the court will subsequently ascertain facts and apply legal norms. It functions as a critical bridge between procedural initiation and substantive adjudication.

Findings of Fact commence with standardized phrasing such as "The Court finds" and present a systematic account of the facts of the case that have been verified through trial-based evidentiary examination and admitted into the record. These findings encompass essential elements, including the origin of the dispute, the sequence of relevant conduct, and the resulting consequences, accompanied by a concise assessment and authentication of the underlying evidence. Beyond reconstructing the objective reality of the case, this section typically concludes by identifying the central issues in dispute and offering preliminary evaluations of the parties' assertions. In doing so, it furnishes a solid factual foundation for the subsequent reasoning, ensuring that the judgment rests upon a verifiable chain of evidence.

Reasoning begins with the phrase "The Court holds" and represents the core analytical component of the judicial decision. Here, the court legally characterizes the established facts, systematically evaluates the principal arguments presented by the parties, and articulates the logical basis for determinations concerning criminal liability, civil responsibility, or the lawfulness of administrative action. Relevant statutory provisions are cited to substantiate the ruling. The section conventionally concludes with a formulaic transition such as "Pursuant to Article [X] of the People's Republic of China [Y] Law, the following judgment is rendered." Through rigorous legal reasoning, this part directly connects factual findings with normative standards, naturally leading to the final disposition and fully demonstrating the rationality, impartiality, and authority of judicial adjudication.

Judgment articulates the court's final and binding ruling in a numbered, itemized format. It possesses immediate legal effect and enforceability, and its language must be precise, unambiguous, and definitive. Depending on the nature of the case, it encompasses determinations of guilt and sentencing, obligations to pay or perform, declarations of rights, or rulings on the validity or modification of administrative acts. As the culmination of the entire judicial decision, the judgment directly responds to the parties' claims and effectuates the judiciary's ultimate allocation of rights and obligations.

Footer contains essential procedural notices and judicial authentication, including advisements regarding the right to appeal and the applicable time frame, cautions concerning liability for delayed performance, signatures of the presiding judges or sole adjudicator and the court clerk, the date of issuance, and the official seal of the people's court. Although devoid of substantive adjudicative content, this portion is indispensable for safeguarding litigants' procedural rights, confirming the document's legal validity, and manifesting judicial transparency and accountability. It is thus an integral element ensuring the completeness and formal integrity of the judicial decision.

B Validation of Qwen-Plus as an Evaluation Tool

Since Qwen-Plus is used to assist the evaluation of ChJGBench, we conduct two validation experiments to examine its reliability under different evaluation settings. The first experiment evaluates the extraction accuracy of Qwen-Plus on objectively verifiable legal elements, where model outputs can be compared against manually checked references. The second experiment examines its ability to perform legal semantic matching, focusing on evaluation dimensions that require meaning-level comparison rather than exact string matching.

B.1 Validation of Qwen-Plus for Extraction Accuracy

To verify the reliability of the Qwen-Plus-based structured extraction process in ChJGBench, we design an extraction accuracy validation experiment. This experiment compares the structured outputs extracted by Qwen-Plus from legal judgments with the corresponding ground-truth results, so as to avoid inaccurate evaluation caused by errors in the extraction stage.

In ChJGBench, different case types, including criminal, civil, and administrative cases, involve multiple extraction dimensions. These dimensions include statutory provisions, disputed issues, criminal charges, monetary fines, recidivism, mitigating circumstances, penalty duration, claim-support decisions, payment obligations, and administrative dispositions. Therefore, it is necessary to validate whether Qwen-Plus can accurately extract these fields from judgment texts.

Data Sampling. We randomly select 1,000 legal judgments from the test set, ensuring that the sampled cases cover criminal, civil, and administrative judgments. Each judgment corresponds to multiple extraction tasks. Based on these judgments, we

construct 5,000 evaluation instances in the format of {*text segment*, *extraction requirement*, *ground truth*}. We then provide these 5,000 instances to Qwen-Plus and require the model to extract the answer according to the given extraction requirement. The model is explicitly instructed not to rewrite or paraphrase the content, but to directly extract the corresponding answer from the original text.

Evaluation Metric. We use different evaluation metrics according to the type of extracted field. For categorical, binary, and numerical fields, we use accuracy as the main metric. Accuracy is defined as:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\hat{y}^{(i)} = y^{(i)}\right), \qquad (6)$$

where $\hat{y}^{(i)}$ denotes the result extracted by Qwen-Plus, $y^{(i)}$ denotes the ground-truth result, and N denotes the number of evaluated instances.

For fields that require normalization, such as monetary amounts, penalty duration, and statutory provisions, we first normalize the extracted results and the ground-truth results into the same format before comparison. The final extraction accuracy reflects the proportion of instances in which Qwen-Plus produces an output exactly consistent with the ground truth.

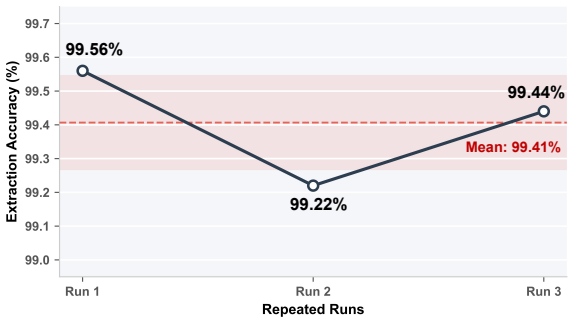

Figure 3: Extraction Accuracy of Qwen-Plus

Conclusion. We set the temperature of Qwen-Plus to 0.3 and repeat the extraction experiment three times to examine the stability of the extraction process. Across the 5,000 extraction instances, Qwen-Plus correctly extracts 4,978, 4,961, and 4,972 instances in the three runs, corresponding to accuracies of 99.56%, 99.22%, and 99.44%, respectively. The average extraction accuracy is 99.41%, indicating that Qwen-Plus achieves highly reliable and stable performance in the structured extraction task.

B.2 Validation of Qwen-Plus for Semantic Matching

The second validation experiment assesses whether Qwen-Plus can judge the legal semantic consistency between generated and reference judgments. This is necessary because several ChJGBench evaluation dimensions, such as disputed issue identification, claim-support decisions, and administrative disposition prediction, require semantic comparison rather than exact string matching. We therefore conduct a semantic matching validation experiment to examine whether Qwen-Plus can reliably identify legal equivalence under constrained evaluation instructions.

Data Sampling. We construct 100 semantic matching instances that cannot be reliably evaluated by regular-expression-based exact matching. Each instance is organized in the following format: {*reference content*, *generated content*, *ground-truth label*}. The sampled instances cover multiple evaluation dimensions that require semantic comparison, including Disputed Issue Identification, Claim Support, and Administrative Disposition Prediction. These dimensions are selected because their correct evaluation often depends on legal meaning rather than surface-level textual overlap.

Human Annotation. We invite three legal professionals to independently annotate the constructed semantic matching instances. For each instance, annotators are asked to judge the relationship between the reference content and the generated content according to the following three labels:

- *Equivalent*: the two contents are legally equivalent, although they may differ in wording. They refer to the same disputed issue, claim-support decision, or administrative disposition.
- *Partially Equivalent*: the two contents partially overlap in legal meaning, but the generated content omits some legally salient elements or contains minor deviations.
- *Not Equivalent*: the two contents correspond to different legal issues, claim outcomes, or administrative dispositions, and therefore cannot be regarded as semantically consistent.

For each instance, we use the majority label assigned by the three human annotators as the ground-truth label.

Qwen-Plus Semantic Matching. We prompt Qwen-Plus to compare the reference content and the generated content for each semantic matching instance. The prompt explicitly instructs the model to ignore differences in wording, expression order, and stylistic variation, and to focus only on legal semantic equivalence. Qwen-Plus is required to output one of the three labels: *Equivalent*, *Partially Equivalent*, or *Not Equivalent*.

Evaluation Metrics. We report Cohen's κ and Krippendorff's α. Cohen's κ measures the agreement between Qwen-Plus and human majority labels while correcting for chance agreement. Krippendorff's α provides another reliability-oriented agreement measure and is suitable for categorical annotation tasks. Since the semantic matching labels are categorical, we compute Krippendorff's α under the nominal distance assumption. Higher values of Cohen's κ and Krippendorff's α indicate stronger consistency between Qwen-Plus and human legal semantic judgments.

Conclusion. The validation results show that Qwen-Plus achieves strong agreement with the human-annotated ground truth, with a Cohen's κ of 0.7727 and a Krippendorff's α of 0.7780. These results indicate that Qwen-Plus can reliably identify legal semantic equivalence under constrained evaluation instructions, even when the generated judgments and reference judgments differ in wording, expression order, or stylistic form. Therefore, Qwen-Plus can serve as a reliable semantic matcher in ChJGBench for evaluation dimensions that cannot be adequately assessed through exact string matching.

C Legal Ability Families

To support a more interpretable analysis beyond individual metrics, we group the evaluation metrics into four legal ability families, as summarized in Table 4. This grouping allows us to examine whether model weaknesses arise from specific isolated metrics or from broader legal capabilities.

D Benchmark Dataset Statistics

Table 5 summarizes the basic statistics of the benchmark dataset. The dataset contains 3,462 Chinese legal judgment samples, with an average prompt length of 2,562.95 characters. Each sample is annotated with a gold answer, supporting reasoning, and judgment result, whose average lengths are 715.48, 595.54, and 119.94 characters, respectively. In addition, each sample cites 4.68 legal provisions on average, indicating that the benchmark requires models not only to generate final judgments but also to ground their predictions in relevant statutory provisions.

Family	Metrics
Legal Basis	Law
Issue & Qualification	Charge, ISS $\mathrm{SUP}_{\mathrm{Adm}}$, $\mathrm{SUP}_{\mathrm{Civ}}$
Judicial Reasoning	JUS, REC MIT, DIS
Disposition & Quantification	Fine_R, Pen_M $\mathrm{PAY_R}_{\mathrm{Adm}}$, $\mathrm{PAY_R}_{\mathrm{Civ}}$

Table 4: Legal ability family taxonomy used for task-family analysis.

Statistic (abbr.)	Value
# Samples	3,462
Lang.	Chinese
Prompt len.	2,562.95 characters
Answer len.	715.48 characters
Reason len.	595.54 characters
Judgment len.	119.94 characters
Cited provisions	4.68

Table 5: Basic statistics of the benchmark dataset. Abbreviations: average prompt length (Prompt len.), average gold answer length (Answer len.), average gold reason length (Reason len.), average gold judgment length (Judgment len.), and average number of cited legal provisions (Cited provisions).

E Case Study

As shown in Figure 4, this case study demonstrates that conventional similarity-based metrics can overestimate consistency in legal judgment evaluation. Although the two texts share highly formulaic legal language, they differ in critical legal dimensions, including the charge, criminal conduct, statutory basis, and involved amount. ROUGE-L and BERTScore assign relatively high scores because they mainly capture lexical overlap and semantic relatedness, while LLM-as-judge provides only a coarse and somewhat subjective assessment. In contrast, our method evaluates key legal dimensions separately and produces a much lower average score, better reflecting the severe legal inconsistency between the two judgments.

Ground Truth		The Court finds that the defendant, Luo Jun, for the purpose of illegal possession, used covert means to enter a residence and ***steal*** property belonging to others valued at RMB 1,010,674, as well as certain property whose value could not be assessed. The amount involved was especially huge, and his conduct constituted the ***crime of theft***. … Accordingly, pursuant to Article ***264***, Paragraph 3 of Article ***67***, and Article 64 of the Criminal Law of the People's Republic of China, as well as Paragraph 1 of Article 1 and Article 14 of the Interpretation of the Supreme People's Court and the Supreme People's Procuratorate on Several Issues Concerning the Application of Law in ***Handling Criminal Cases*** of Theft…
Candidate Answer		The Court finds that the defendant, Luo Jun, for the purpose of illegal possession, used such means as concealing facts and fabricating reasons to ***defraud*** others of property valued at RMB 1 million. The amount involved was especially huge, and his conduct constituted the ***crime of fraud***. … Accordingly, pursuant to Article ***266*** and Article ***67*** of the Criminal Law of the People's Republic of China, as well as the Interpretation of the Supreme People's Court and the Supreme People's Procuratorate on Several Issues Concerning the Application of Law in Handling Property Crimes ***Involving Home Entry***…
Different Evaluation Metrics	**Rouge-L**	0.6986 ❌ **Formulaic judgment language creates long overlapping sequences, so ROUGE-L can score highly despite key legal fact differences.**
	BertScore	0.9058 ❌ **BERTScore captures semantic relatedness, not factual consistency, so critical local errors may be hidden by overall similarity.**
	LLM-as-judge	3/10 ⚠️ **Coarse-grained scoring and potential subjectivity make the evaluation less stable.**
	Ours	Law = 0.25, Charge = 0, ISS = 0.1, JUS = 0.15 ✅ >>> **Avg. = 0.125** 🏆 **More fine-grained and legally grounded.**

Figure 4: Case study comparing different evaluation metrics on a legally inconsistent judgment pair.