

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

Yinghao Hu¹, Linggang Kong², Qian Xiao³, Yuting Huang¹, Yunlong Zhang²,
Yujie Zhang¹, Fei Duan³, Yiquan Wu^{1,*}, Kaisong Song^{3,4},
Jun Lin³, Kun Kuang^{1,*}, Fei Wu¹

¹Zhejiang University, Hangzhou, China

²National University of Defense Technology, Changsha, China

³Alibaba Group, Hangzhou, China

⁴Northeastern University, Shenyang, China

Emails: {huyinghao, wuyiquan, kunkuang}@zju.edu.cn

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 generating the reasoning and dispositive sections of Chinese judgments across civil, criminal, and administrative cases. Built from 3,462 real-world judgments, ChJGBench provides a case-type-aware diagnostic evaluation framework comprising 13 metric types and adopts a hybrid paradigm. Experiments reveal no uniform advantage for legal-domain models over general-purpose models, with model strengths varying across diagnostic dimensions. Performance is generally stronger on criminal cases, while civil and administrative cases continue to present substantial challenges.

1 Introduction

In China, written judgments are the primary official records through which courts present factual findings, legal reasoning, and dispositive outcomes. Generating their core adjudicative sections requires models not only to produce fluent legal text, but also to resolve outcome-critical elements such as claims, charges, penalties, monetary obligations, and statutory grounds. Prior work has studied criminal court-view generation from fact descriptions (Ye et al., 2018; Liu et al., 2024; Xu et al., 2024). However, existing evaluations still have limitations in coverage and granularity:

Benchmark	M-Type	Samp.	Dims.	Eval. Method
JuDGE	×	501	4	Det.
CaseGen	×	500	15	LLM
J1-EVAL	×	162	6	Det. + LLM
PLawBench	N/A	40	6	Rubric + LLM
ChJGBench	✓	3,462	13	Diag. + LLM

Table 1: Comparison of **ChJGBench** with prior legal judgment and document generation benchmarks. **M-Type**: all three litigation types; **Samp.**: distinct source cases; **Dims.**: dimensions for document drafting and judgment generation; **Det.**: deterministic evaluation, e.g., exact, pattern, or regex matching; **Diag.**: diagnostic evaluation with litigation-type-specific criteria.

studies of court-view generation largely focus on criminal cases, while multi-type benchmarks use smaller datasets and comparatively coarse or heterogeneous evaluation criteria. Consequently, it remains unclear how reliably current LLMs generate judicial reasoning and dispositive outcomes across litigation types under a unified, fine-grained diagnostic framework.

Existing legal benchmarks (Fei et al., 2024; 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 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 include fewer samples and evaluation dimensions, or cover only a subset of litigation types. Moreover, commonly used text-

*Corresponding authors.

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	Monetary Fine	4.4
REC	Recidivism Assessment	4.4
MIT	Mitigating Circumstances	4.4
Pen_M	Penalty Duration Prediction	4.4
SUP_{Civ}	Claim Support	4.5
PAY_{Civ}	Payment Obligation	4.5
SUP_{Adm}	Claim Support	4.6
PAY_{Adm}	Payment Obligation	4.6
DIS	Administrative Disposition	4.6

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

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.

To address these limitations, we introduce **ChJGBench**, a comprehensive benchmark for 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 shared and litigation-type-specific dimensions. The shared dimensions assess statutory provision application, disputed issue identification, and judicial justifica-

tion. The case-specific dimensions capture distinctive adjudicative requirements in different litigation types: criminal cases involve charge prediction, monetary fine, recidivism assessment, mitigating circumstance identification, and penalty duration prediction; civil cases involve claim support and payment obligation; and administrative cases involve claim support, administrative disposition, and payment obligation. In total, ChJGBench defines 13 distinct metric types: three shared metrics and ten litigation-specific metrics. These correspond to 19 case-type-specific metric instances in total.

To improve evaluation reliability and scalability, ChJGBench adopts a hybrid metric computation paradigm. LLMs are used in a constrained manner for legal-element extraction and normalization. For ISS, JUS, and DIS, an evaluator LLM additionally performs pairwise semantic tagging under dimension-specific rubrics. The resulting structured elements and semantic tags are then converted into scores using deterministic, dimension-specific scoring procedures. In this way, LLMs serve as controlled information extractors and semantic taggers rather than unconstrained subjective judges. The LLM-based extraction prompts for all evaluation dimensions are provided in Appendix I, while the semantic-tagging prompts for ISS, JUS, and DIS are provided in Appendix J.

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 adoption of LLMs in the legal domain (Deng et al., 2023; Shui et al., 2023; Louis et al., 2024), 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., 2024) 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

Prior work on judicial judgment generation has primarily focused on criminal cases. Early studies explore court-view generation from fact descriptions (Ye et al., 2018), while subsequent work decomposes the task according to legal outcomes such as charges, law articles, and prison terms (Liu et al., 2024), or legal concepts such as recidivism and confession (Xu et al., 2024). Evaluation in these settings often relies in part on general text-similarity metrics, such as ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020). While useful for measuring textual or semantic similarity, such metrics provide limited information about legally consequential distinctions, such as charge determination and sentencing outcomes. More recent work has moved toward broader and more structured evaluation. SimuCourt (He et al., 2024) covers 420 criminal, civil, and administrative judgments and combines deterministic, LLM-based, and human evaluation. JuDGE (Su et al., 2025) provides more fine-grained evaluation of charge, penalty, and legal-reference accuracy, although it focuses on criminal judgments. CaseGen (Li et al., 2025), J1-EVAL (Jia et al., 2026), and PLawBench (Shi et al., 2026) further employ LLM-based or rubric-based evaluation for broader legal generation and

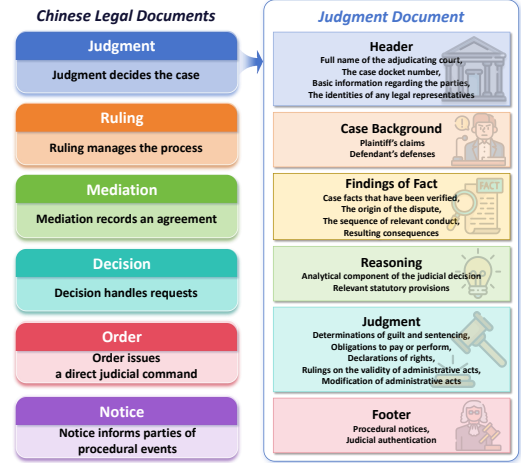


Figure 1: Structure of a legal judgment document.

drafting tasks.

These studies substantially expand the scope and granularity of legal generation evaluation, but a unified framework for fine-grained evaluation of judicial reasoning and dispositive outcomes across criminal, civil, and administrative judgments remains relatively underexplored. ChJGBench aims to contribute in this direction by evaluating these three litigation types within a unified framework that combines deterministic matching with LLM-based semantic evaluation.

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.

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

We use *Header*, *Background*, and *Findings of Fact* as the input because they provide the procedural and factual context necessary for adjudication. We exclude *Footer* because it is largely formulaic.

The generation target comprises *Reasoning* and *Judgment*, as these two 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.

In summary, the legal judgment generation task takes $\mathcal{X}$ as input and produces $\mathcal{Y}$ as output, where

$$\begin{aligned}\mathcal{X} &= (\text{Header}, \text{Background}, \text{Findings of Fact}) \\ \mathcal{Y} &= (\text{Reasoning}, \text{Judgment}).\end{aligned}\quad (2)$$

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 Online¹, 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:

$$\tilde{P}_\tau(c | t, y) = \frac{P_{t,y}(c)^\tau}{\sum_k P_{t,y}(k)^\tau}, \quad \tau \in (0, 1]. \quad (3)$$

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

¹<https://wenshu.court.gov.cn/>

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

$$Q_{\lambda,\tau}(c | t, y) = \lambda P_{t,y}(c) + (1 - \lambda) \tilde{P}_\tau(c | t, y), \quad (4)$$

where $\lambda \in [0, 1]$ balances empirical fidelity and long-tail coverage. For each case type, instances are sampled according to $Q_{\lambda,\tau}$ until the target sample size is reached. We set $\tau = 0.5$ and $\lambda = 0.3$ as the sampling hyperparameters; Appendix B provides a detailed discussion of their effects.

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 G.

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, can introduce subjectivity and instability, especially for long and complex judgments.

As summarized in Table 2, the framework does not apply a uniform metric set to all judgments. It organizes the 13 metric types into three shared dimensions and ten litigation-specific dimensions. The shared dimensions capture requirements common to civil, criminal, and administrative cases, while the litigation-specific dimensions reflect the distinctive adjudicative tasks of each litigation type.

Within this structure, the scoring pipeline separates constrained LLM-based processing from deterministic score computation. Under explicit prompts, LLMs extract or normalize predefined legal elements from generated judgments. For ISS, JUS, and DIS, an evaluator LLM performs pairwise semantic tagging under dimension-specific rubrics. The resulting elements and tags are converted into scores using deterministic procedures tailored to each dimension: set-based F_1 for Law and Charge; macro-averaged F_1 for REC, MIT, and SUP; NVS for Fine and PAY; Pen_M scoring based on NVS for numerical durations and exact matching for categorical punishments; and soft semantic F_1 aggregation for ISS, JUS, and DIS.

By restricting LLMs to controlled extraction, normalization, and rubric-guided semantic tagging, the framework avoids unconstrained holistic scoring. The resulting evaluation is both diagnostic and legally grounded: it assesses whether generated judgments reach correct conclusions on key adjudicative components. Appendix C further specifies, for each dimension, the LLM processing components and final scoring method.

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 the reference provisions. It is applicable to civil, criminal, and administrative cases. Each provision is represented as [Statute Name] Article X. Statute names are standardized using a predefined alias set, and each combination of a statute name and an article number is treated as one item. We compute an F_1 score for each case by exact matching and report the mean across cases.

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 use an LLM to extract disputed-issue lists from the judgment generated by the model under evaluation and the corresponding real case. Given the source legal text, the model-output extract list, and the ground-truth content list, an evaluator LLM assigns exactly one semantic tag to every possible pair consisting of one ground-

truth content item and one model-output extract under the ISS-specific rubric, and the case-level score is computed deterministically from these tags using Equation 36.

Judicial Justification (JUS). This metric evaluates whether the generated judgment provides the legally material analysis and grounds explaining why the court reaches its disposition, including how adjudicated facts and applicable law support its conclusions on disputed issues, claims, or charges. It measures how well the generated reasoning covers and matches the ground-truth fact-rule-conclusion points, rather than evaluating linguistic fluency or overall textual similarity. We first use an LLM to extract lists of key reasoning points from the judgment generated by the model under evaluation and the corresponding real case. Given the source legal text, the model-output extract list, and the ground-truth content list, an evaluator LLM assigns exactly one semantic tag to every possible pair consisting of one ground-truth content item and one model-output extract according to the JUS-specific rubric, and the case-level score is computed deterministically from these tags using Equation 36.

4.4 Criminal Case Dimensions

Criminal Charge Prediction (Charge). This metric evaluates criminal charge prediction by comparing the charges extracted from the generated judgment with the ground-truth charges for each defendant. We use an LLM to extract the predicted charge labels and compute the F_1 score through label matching.

Monetary Fine (Fine). This metric measures the numerical accuracy of monetary fines imposed in criminal cases. For a predicted amount a and the corresponding reference amount b , we compute Normalized Value Similarity (NVS) as

$$\text{NVS}(a, b) = 1 - \frac{|a - b|}{\max(a, b)}, \quad (5)$$

where $a, b \geq 0$ and $a + b > 0$. For cases with multiple defendants, we compute the case-level Fine score by averaging the defendant-level NVS scores over all defendants who receive a fine in the reference judgment. A defendant receives a score of 0 if the corresponding fine amount is missing from the prediction. The final Fine score is the mean of the resulting case-level scores.

Recidivism Assessment (REC). This metric evaluates whether the generated judgment correctly identifies each 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 and report the macro-averaged F_1 score over the recidivist and non-recidivist classes, with each defendant treated as one instance.

Mitigating-Circumstance Identification (MIT). This metric evaluates whether the generated judgment correctly identifies the presence of a mitigating circumstance for each defendant. Such circumstances may include voluntary surrender, confession, remorse, restitution, compensation, or obtaining the victim’s forgiveness. We formulate this dimension as a defendant-level binary classification task and report the macro-averaged F_1 score over the positive and negative classes.

Penalty Duration Prediction (Pen_M). This metric evaluates whether the generated judgment predicts the correct principal punishment. It covers duration-based punishments, including fixed-term imprisonment, criminal detention, and public surveillance, as well as life imprisonment and the death penalty. Let $\mathcal{T} \subset \mathbb{R}_{>0}$ denote punishment durations normalized to months, and let $\mathcal{C}$ contain the categorical labels LIFE_IMPRISONMENT and DEATH_PENALTY. For each defendant with a reference punishment, let $\hat{p}_i$ and p_i denote the predicted and reference punishments, respectively. The defendant-level score is defined as

$$s_i^{\text{Pen_M}} = \begin{cases} \text{NVS}(\hat{p}_i, p_i), & \hat{p}_i, p_i \in \mathcal{T}, \\ 1, & \hat{p}_i = p_i \in \mathcal{C}, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

For each case, the case-level Pen_M score is the mean of the defendant-level scores for all defendants with an applicable reference punishment. The final Pen_M score is the mean of the case-level scores over cases containing at least one such defendant.

4.5 Civil Case Dimensions

Claim Support (SUP_{Civ}). This metric evaluates whether the generated civil judgment correctly predicts the overall disposition of the plaintiff’s claims as upheld, partially upheld, or rejected. It captures

the substantive adjudicative outcome of civil litigation. We formulate it as a three-class classification task and report macro- F_1 over the three outcome labels.

Payment Obligation (PAY_{Civ}). This metric evaluates whether the generated civil judgment accurately predicts the total amount of payment obligations imposed on each obligor, including compensation, repayment, damages, restitution, and other monetary liabilities. We evaluate obligor-level numerical accuracy using NVS, as defined in Equation 5, and average the resulting scores first within each case and then across cases. Evaluation is restricted to civil cases containing at least one reference payment obligation.

4.6 Administrative Case Dimensions

Claim Support (SUP_{Adm}). This metric evaluates whether the generated administrative judgment correctly predicts the overall disposition 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. The disposition is classified as upheld, partially upheld, or rejected, and performance is measured using macro- F_1 over the three outcome labels.

Payment Obligation (PAY_{Adm}). This metric evaluates whether the generated administrative judgment accurately predicts the amounts of monetary obligations, such as administrative compensation, restitution, refunds, or other payment-related remedies. We compute PAY_{Adm} in the same way as PAY_{Civ}.

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 use an LLM to extract lists of atomic core dispositions from the judgment generated by the model under evaluation and the corresponding real case, retaining each disposition’s operative action, object, and any explicitly stated scope or condition. Given the source legal text, the model-output extract list, and the ground-truth content list, an evaluator LLM assigns exactly one semantic tag to every possible

pair consisting of one ground-truth content item and one model-output extract, and the case-level score is computed deterministically using Equation 36. Evaluation is restricted to administrative cases with explicit dispositive language in the corresponding real case.

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. The experiments examine two categories of LLMs on legal judgment generation tasks: legal-domain LLMs and mainstream general-purpose LLMs. The two legal-domain LLMs are WisdomInterrogatory (Wu et al., 2026) and LegalOne (Li et al., 2026). The seven mainstream general-purpose LLMs are 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 Qwen3-Max (Yang et al., 2025).

Metrics. We evaluate model outputs using the case-type-aware metric framework described above. Table 2 summarizes the evaluation dimensions, and Appendix Table 5 lists their scoring methods. All scores are reported to three decimal places.

Evaluation Model Validation. To assess model outputs in a consistent manner, we use Qwen3-Plus as the evaluation model for information extraction and necessary semantic matching. The extraction accuracy of Qwen3-Plus and its agreement with human annotations in semantic matching are reported in Appendix D.1 and Appendix D.2, respectively. Specifically, Qwen3-Plus achieves an average extraction accuracy of 99.41%, demonstrating reliable and stable structured extraction performance. For the three tasks that require semantic matching, Qwen3-Plus also shows strong agreement with human annotations, with Cohen’s κ values of 0.7036 for JUS, 0.8203 for ISS, and 0.8909 for DIS.

Experimental Details. We use few-shot prompting for all tasks, with the detailed generation prompt provided in Appendix H. To ensure fairness and comparability, all models use the same prompt template, demonstration examples, input context, and output constraints. For judgment generation in the main benchmark experiments, the decoding temperature is set to 0 whenever temperature control is supported.

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 pattern suggests that training small-scale models on legal-domain data may help incorporate 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, while Grok-4 shows strong performance on criminal ISS and Charge, as well as administrative DIS. Claude Sonnet 4.5 performs best on civil and administrative JUS, and Qwen3-Max obtains the highest score on civil ISS. This pattern suggests that different models capture different aspects of legal judgment generation. Some models show stronger capabilities in JUS and ISS, whereas others are more effective in predicting concrete legal outcomes.

Criminal cases show relatively stronger performance, whereas civil and administrative cases remain comparatively challenging. Focusing on Law, JUS, and ISS, the three dimensions shared across all litigation types, we observe a descriptive pattern: models generally achieve higher average scores on criminal cases, followed by administrative cases, while obtaining lower scores on civil cases. One possible explanation is that criminal judgments may be relatively more homogeneous in structure and reasoning patterns, which may make the mapping from established facts to legal conclusions more consistent across cases. In contrast,

Case / Metric	Model								
	WisdomI.	LegalOne	Claude Sonnet 4.5	DS-R1	Gemini 3 Pro	GPT-5.4	Grok-4	LLaMA 405B	Qwen3 Max
Criminal									
Law	0.393 $\pm$ 0.006	<u>0.718</u> $\pm$ 0.007	0.677 $\pm$ 0.003	0.687 $\pm$ 0.004	0.746 $\pm$ 0.002	0.642 $\pm$ 0.003	0.610 $\pm$ 0.002	0.504 $\pm$ 0.006	0.643 $\pm$ 0.003
JUS	0.369 $\pm$ 0.025	0.714 $\pm$ 0.002	0.734 $\pm$ 0.009	0.758 $\pm$ 0.002	0.738 $\pm$ 0.001	0.702 $\pm$ 0.001	<u>0.739</u> $\pm$ 0.002	0.639 $\pm$ 0.003	0.734 $\pm$ 0.005
ISS	0.493 $\pm$ 0.020	0.802 $\pm$ 0.003	0.812 $\pm$ 0.005	0.814 $\pm$ 0.006	<u>0.823</u> $\pm$ 0.000	0.786 $\pm$ 0.002	0.824 $\pm$ 0.002	0.702 $\pm$ 0.010	0.797 $\pm$ 0.003
Charge	0.468 $\pm$ 0.028	0.973 $\pm$ 0.000	<u>0.974</u> $\pm$ 0.003	0.970 $\pm$ 0.004	<u>0.974</u> $\pm$ 0.003	0.973 $\pm$ 0.002	0.977 $\pm$ 0.001	0.960 $\pm$ 0.004	0.972 $\pm$ 0.004
Fine	0.134 $\pm$ 0.001	0.564 $\pm$ 0.001	<u>0.577</u> $\pm$ 0.001	0.562 $\pm$ 0.001	0.611 $\pm$ 0.001	0.574 $\pm$ 0.001	0.502 $\pm$ 0.002	0.476 $\pm$ 0.002	<u>0.594</u> $\pm$ 0.001
REC	0.108 $\pm$ 0.000	<u>0.736</u> $\pm$ 0.007	0.704 $\pm$ 0.001	0.713 $\pm$ 0.005	0.753 $\pm$ 0.002	0.717 $\pm$ 0.005	0.667 $\pm$ 0.007	0.496 $\pm$ 0.017	<u>0.705</u> $\pm$ 0.012
MIT	0.365 $\pm$ 0.038	0.967 $\pm$ 0.000	0.966 $\pm$ 0.002	0.960 $\pm$ 0.002	0.974 $\pm$ 0.001	0.963 $\pm$ 0.001	<u>0.971</u> $\pm$ 0.003	0.882 $\pm$ 0.003	0.970 $\pm$ 0.002
Pen_M	0.604 $\pm$ 0.003	0.731 $\pm$ 0.000	0.750 $\pm$ 0.001	0.740 $\pm$ 0.000	0.786 $\pm$ 0.000	0.748 $\pm$ 0.001	0.724 $\pm$ 0.001	0.602 $\pm$ 0.001	<u>0.760</u> $\pm$ 0.004
Macro Avg	0.367 $\pm$ 0.007	<u>0.776</u> $\pm$ 0.000	0.774 $\pm$ 0.002	<u>0.776</u> $\pm$ 0.001	0.801 $\pm$ 0.001	0.763 $\pm$ 0.001	0.752 $\pm$ 0.001	0.658 $\pm$ 0.002	<u>0.772</u> $\pm$ 0.002
Civil									
Law	0.079 $\pm$ 0.013	0.424 $\pm$ 0.003	0.442 $\pm$ 0.002	0.232 $\pm$ 0.006	0.595 $\pm$ 0.002	<u>0.465</u> $\pm$ 0.003	0.397 $\pm$ 0.006	0.173 $\pm$ 0.003	0.216 $\pm$ 0.002
JUS	0.402 $\pm$ 0.043	0.590 $\pm$ 0.005	0.722 $\pm$ 0.003	0.698 $\pm$ 0.001	0.657 $\pm$ 0.003	0.635 $\pm$ 0.003	<u>0.712</u> $\pm$ 0.003	0.511 $\pm$ 0.017	0.650 $\pm$ 0.004
ISS	0.622 $\pm$ 0.029	0.741 $\pm$ 0.001	0.814 $\pm$ 0.002	<u>0.821</u> $\pm$ 0.002	0.812 $\pm$ 0.002	0.761 $\pm$ 0.003	0.818 $\pm$ 0.005	0.745 $\pm$ 0.009	0.826 $\pm$ 0.002
SUP _{Civ}	0.626 $\pm$ 0.029	0.786 $\pm$ 0.001	0.780 $\pm$ 0.004	0.782 $\pm$ 0.004	0.828 $\pm$ 0.005	<u>0.800</u> $\pm$ 0.003	0.760 $\pm$ 0.001	0.685 $\pm$ 0.006	0.741 $\pm$ 0.003
PAY _{Civ}	0.245 $\pm$ 0.004	0.744 $\pm$ 0.004	<u>0.747</u> $\pm$ 0.004	<u>0.756</u> $\pm$ 0.004	0.758 $\pm$ 0.003	0.747 $\pm$ 0.004	0.711 $\pm$ 0.004	0.617 $\pm$ 0.016	0.734 $\pm$ 0.002
Macro Avg	0.395 $\pm$ 0.023	0.657 $\pm$ 0.002	<u>0.701</u> $\pm$ 0.001	0.658 $\pm$ 0.001	0.730 $\pm$ 0.001	0.682 $\pm$ 0.002	0.680 $\pm$ 0.003	0.546 $\pm$ 0.008	0.634 $\pm$ 0.001
Administrative									
Law	0.115 $\pm$ 0.024	0.379 $\pm$ 0.004	<u>0.563</u> $\pm$ 0.005	0.542 $\pm$ 0.004	0.578 $\pm$ 0.005	0.516 $\pm$ 0.002	0.544 $\pm$ 0.005	0.257 $\pm$ 0.009	0.429 $\pm$ 0.003
JUS	0.305 $\pm$ 0.019	0.597 $\pm$ 0.007	0.735 $\pm$ 0.005	0.716 $\pm$ 0.003	0.668 $\pm$ 0.003	0.638 $\pm$ 0.006	<u>0.731</u> $\pm$ 0.007	0.520 $\pm$ 0.007	0.689 $\pm$ 0.001
ISS	0.532 $\pm$ 0.018	0.724 $\pm$ 0.002	0.798 $\pm$ 0.004	0.807 $\pm$ 0.002	0.789 $\pm$ 0.002	0.745 $\pm$ 0.003	0.795 $\pm$ 0.003	0.691 $\pm$ 0.002	<u>0.801</u> $\pm$ 0.004
SUP _{Adm}	0.685 $\pm$ 0.006	0.808 $\pm$ 0.008	0.825 $\pm$ 0.002	<u>0.841</u> $\pm$ 0.006	0.852 $\pm$ 0.003	0.826 $\pm$ 0.002	0.825 $\pm$ 0.002	0.744 $\pm$ 0.001	0.829 $\pm$ 0.005
PAY _{Adm}	0.037 $\pm$ 0.003	0.351 $\pm$ 0.005	0.368 $\pm$ 0.006	0.222 $\pm$ 0.005	0.441 $\pm$ 0.001	<u>0.429</u> $\pm$ 0.004	0.307 $\pm$ 0.005	0.227 $\pm$ 0.004	0.311 $\pm$ 0.003
DIS	0.332 $\pm$ 0.020	0.574 $\pm$ 0.008	0.765 $\pm$ 0.001	<u>0.787</u> $\pm$ 0.003	0.732 $\pm$ 0.004	<u>0.701</u> $\pm$ 0.005	0.796 $\pm$ 0.004	0.713 $\pm$ 0.003	0.685 $\pm$ 0.008
Macro Avg	0.334 $\pm$ 0.015	0.572 $\pm$ 0.003	<u>0.676</u> $\pm$ 0.003	0.653 $\pm$ 0.003	0.677 $\pm$ 0.003	0.642 $\pm$ 0.002	0.666 $\pm$ 0.001	0.525 $\pm$ 0.001	0.624 $\pm$ 0.003

Table 3: Model performance across evaluation dimensions and case types. Values are reported as mean $\pm$ sample standard deviation over three runs. Macro Avg is computed per run as the unweighted mean of the unrounded component scores within each case type. Values are rounded to three decimals. Bold and underlined values indicate the best and second-best means, respectively; means that round to the same three-decimal value share the same rank. Abbreviations: WisdomI. = WisdomInterrogatory; DS-R1 = DeepSeek-R1.

civil and administrative judgments may involve more heterogeneous legal relationships, claims, procedural questions, and remedies. These differences may partly account for the observed variation in model performance across litigation types.

5.3 Analysis of Task Difficulty and Cross-Model Variability

As shown in Figure 2, each point represents a case-type-specific task. The horizontal coordinate is the unweighted mean of the nine model-level mean scores reported in Table 3, while the vertical coordinate is their sample standard deviation. The dashed vertical and horizontal lines mark the respective medians of these quantities across the 19 tasks. Point colors indicate the corresponding legal ability families, whose taxonomy is provided in Appendix E. The error bars show how variability across the three runs propagates to the task-level mean and cross-model standard deviation; their calculation is detailed in Appendix F. For visual clarity, error-bar half-widths below 0.005 are omit-

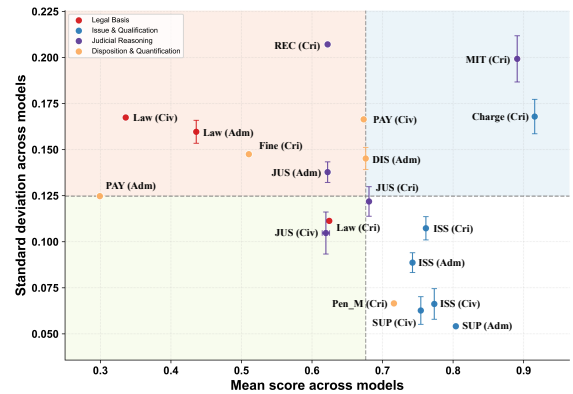


Figure 2: Task difficulty and cross-model variability across evaluation dimensions.

ted independently for each axis.

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 (Adm), Fine (Cri), and REC (Cri) also appear on the left side, indicating that administrative payment-amount prediction, criminal fine-amount prediction, and recidivism assessment remain challenging tasks.

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

In contrast, the lower-right quadrant includes dimensions with high mean scores and low cross-model variability, 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 cross-model variability makes them **less effective at differentiating the capabilities** of advanced models.

6 Conclusions

In this paper, we introduce **ChJGBench**, a comprehensive and diagnostically grounded benchmark for Chinese legal judgment generation. ChJGBench is built from 3,462 diverse real-world judgments across civil, criminal, and administrative cases and 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: models achieve higher overall scores on criminal cases but perform less strongly on civil and administrative cases. These results highlight the need to improve models' ability to address the diverse legal relationships, claims, remedies, and statutory requirements involved in civil and administrative adjudication.

Limitations

ChJGBench has several limitations. First, because it is constructed from publicly available court judgments, it may not fully represent cases that are unpublished or otherwise unavailable, potentially introducing selection bias. Second, the current version of ChJGBench constitutes a static snapshot, whereas statutes, judicial interpretations, and adjudicatory standards may evolve over time. To partially address this limitation, the reusable construction and stratified sampling pipeline proposed in this paper can be rerun periodically to incorporate new judgments, emerging case categories, and relevant legal developments, thereby enabling targeted and versioned updates. Third, the use of publicly available judgments creates a potential risk of data contamination. Although we remove duplicate documents and apply structural and quality controls, we cannot fully determine whether benchmark cases, closely related cases, or derivative materials were included in the training data of the evaluated models, particularly for closed-source models.

Ethical Statement

ChJGBench is constructed from publicly available court judgments collected from China Judgments Online, the official national repository of Chinese court documents. Although the source documents are publicly accessible, judicial records may contain sensitive or potentially identifying information. As part of the release process, we remove or mask personally identifiable and reverse-searchable information and conduct an additional privacy review prior to any public release to reduce privacy and re-identification risks.

ChJGBench is intended solely for research on the capabilities and limitations of language models in legal judgment generation. It must not be used

to support or justify real-world judicial decision-making without rigorous human oversight, legal validation, institutional accountability, and compliance with applicable laws and regulations. Benchmark performance should not be interpreted as evidence of deployment readiness or as a substitute for professional legal judgment. We also acknowledge that judicial data may reflect historical, institutional, or social biases, which models trained on or evaluated against such data may reproduce.

Acknowledgements

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02037), Research Project of the Philosophy and Social Sciences Planning Research Platform (Laboratory) of Zhejiang Province, and Key R&D Program of Hangzhou (2025SZDA0254). It was also supported by the National Natural Science Foundation of China (62106039) and Alibaba Research Intern Program. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding organizations.

We declare that AI-based writing tools were used only for limited language polishing and grammatical refinement in preparing this paper. The authors have carefully verified all AI-assisted text and take full responsibility for the content of the paper.

References

- Anthropic. 2025. [Claude Sonnet 4.5 System Card](#).
- Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. [LexGLUE: A benchmark dataset for legal language understanding in English](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.
- Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. [Syllogistic reasoning for legal judgment analysis](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009, Singapore. Association for Computational Linguistics.
- 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 *International Conference on Learning Representations*, volume 2026, pages 43451–43489.
- 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.
- 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, et al. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- 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.
- 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.
- Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. 2024. [AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9399–9416, Miami, Florida, USA. Association for Computational Linguistics.
- 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](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10351–10376, San Diego, California, United States. Association for Computational Linguistics.

- 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](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 25061–25094. Curran Associates, Inc.
- 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, Barcelona, Spain. Association for Computational Linguistics.
- Yifei Liu, Yiquan Wu, Ang Li, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2024. [Unleashing the power of LLMs in court view generation by stimulating internal knowledge and incorporating external knowledge](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2782–2792, Mexico City, Mexico. Association for Computational Linguistics.
- Antoine Louis, Gijs Van Dijck, and Gerasimos Spanakis. 2024. [Interpretable long-form legal question answering with retrieval-augmented large language models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22266–22275.
- 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Yuzhen Shi, Huanghai Liu, Yiran HU, Song Gaojie, Xu Xinran, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Feng Di, Wenbo Lv, Weiheng Wu, Kexin Yang, Sen Yang, Wei Wang, Rongyao Shi, Qiu Yuanyang, Yuemeng Qi, Zhang Jingwen, and 11 others. 2026. [PLAWBENCH: A rubric-based benchmark for evaluating LLMs in real-world legal practice](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10067–10116, San Diego, California, United States. Association for Computational Linguistics.
- 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*, pages 3573–3583, New York, NY, USA. Association for Computing Machinery.
- 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.
- Qi Xu, Xiao Wei, Hang Yu, Qian Liu, and Hao Fei. 2024. [Divide and conquer: Legal concept-guided criminal court view generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3395–3410, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. [Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions](#). 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 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

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.

B Hyperparameters for Benchmark Data Distribution

Hyperparameter Settings. The hyperparameters used in our coverage-oriented sampling strategy for constructing the benchmark data distribution are summarized in Table 4.

Trade-off Interpretation. More generally, the formulation provides a tunable trade-off between preserving the empirical category distribution and shifting the sampling distribution toward uniformity across legal categories. When either $\tau = 1$ or $\lambda = 1$, the resulting sampling distribution is identical to the empirical distribution. Conversely, when $\lambda = 0$ and τ approaches zero from above, the sampling distribution approaches a uniform distribution over the observed categories within each litigation-type-and-year stratum.

C Evaluation Components and Scoring Methods

Table 5 provides a compact summary of all evaluation dimensions, indicating whether each dimension involves (1) LLM-based extraction or normalization of structured legal elements, (2) LLM-based semantic matching between the generated and reference judgments, and (3) the final scoring method.

D Validation of Qwen3-Plus as an Evaluation Tool

Since Qwen3-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 Qwen3-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.

D.1 Validation of Qwen3-Plus for Extraction Accuracy

In ChJGBench, Qwen3-Plus is used to extract or normalize predefined legal elements from generated judgments. To provide a supporting empirical check of this process, we conduct an extraction accuracy validation experiment on a sample of manually verifiable extraction instances.

Data Sampling. We randomly sample 1,000 cases from the full benchmark evaluation set to validate the extraction capability of Qwen3-Plus. Each sampled case can be used to construct multiple extraction tasks. Ultimately, we construct 5,000 evaluation instances in the format of *text segment*, *extraction requirement*, *ground truth*. We then provide these 5,000 instances to Qwen3-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 instance-level accuracy as the validation metric. An extraction is counted as correct when it satisfies the corresponding extraction requirement and matches the manually checked ground truth. Accuracy is defined as:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}^{(i)} = y^{(i)}), \quad (7)$$

where $\hat{y}^{(i)}$ denotes the result extracted by Qwen3-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 Qwen3-Plus produces an output exactly consistent with the ground truth.

Conclusion. We repeat the extraction experiment three times to examine the stability of the extraction process. Across the 5,000 extraction instances,

Hyperparameter	Setting	Function
Power-temperature coefficient τ	0.5	Controls the degree of smoothing applied to the empirical category distribution.
Interpolation coefficient λ	0.3	Controls the balance between the empirical distribution and the temperature-smoothed distribution.

Table 4: Hyperparameter settings used in the coverage-oriented sampling strategy. Setting $\tau = 0.5$ applies a square-root transformation to the empirical category distribution, reducing the relative dominance of high-frequency categories and increasing the relative sampling probabilities of low-frequency categories. With $\lambda = 0.3$, the final sampling distribution combines 30% of the empirical distribution with 70% of the temperature-smoothed distribution, thereby emphasizing long-tail coverage while retaining the overall structure of the empirical distribution.

Dimension	LLM Extraction	LLM Semantic Matching	Final Scoring Method	Main-text Section
Law	✓	×	Case-level set F_1	Sec. 4.3
ISS	✓	✓	Soft semantic F_1	Sec. 4.3
JUS	✓	✓	Soft semantic F_1	Sec. 4.3
Charge	✓	×	Defendant-level set F_1	Sec. 4.4
Fine	✓	×	NVS	Sec. 4.4
REC	✓	×	Binary macro- F_1	Sec. 4.4
MIT	✓	×	Binary macro- F_1	Sec. 4.4
Pen_M	✓	×	NVS & exact match	Sec. 4.4
SUP_{Civ}	✓	×	Three-class macro- F_1	Sec. 4.5
PAY_{Civ}	✓	×	NVS	Sec. 4.5
SUP_{Adm}	✓	×	Three-class macro- F_1	Sec. 4.6
PAY_{Adm}	✓	×	NVS	Sec. 4.6
DIS	✓	✓	Soft semantic F_1	Sec. 4.6

Table 5: Evaluation components and scoring methods for each ChJGBench evaluation dimension. LLM extraction denotes the structured extraction or normalization of predefined legal elements. LLM semantic matching denotes constrained element-level comparison between the generated and reference judgments.

Qwen3-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%, providing supporting evidence that Qwen3-Plus performs reliably on the sampled extraction instances.

D.2 Dimension-Specific Validation of Qwen3-Plus for Semantic Matching

This validation experiment assesses whether Qwen3-Plus can reliably assign semantic-consistency labels for ISS, JUS, and DIS. We conduct separate agreement analyses for the three dimensions by comparing the labels assigned by Qwen3-Plus with those provided by human annotators.

Data Sampling. For each of ISS, JUS, and DIS, we randomly sample 50 validation instances, yielding 150 instances in total. Each instance consists of a triplet containing the *source legal text*, a *model-*

output extract, and the corresponding *ground-truth content*. The model-output extract is extracted by Qwen3-Plus from a judgment generated by the model under evaluation, while the ground-truth content is extracted by Qwen3-Plus from the corresponding real case. Given this triplet, Qwen3-Plus and human annotators independently assess the legal-semantic consistency between the model-output extract and the ground-truth content, using the source legal text as case context.

Agreement Labels. We ask Qwen3-Plus and three legal professionals to independently annotate the constructed validation instances. For each validation triplet, the annotators use the source legal text as case context and assess the legal-semantic consistency between the model-output extract and the ground-truth content by assigning one of the following three labels:

- **Equivalent:** The model-output extract and the ground-truth content are legally equivalent despite possible differences in wording,

Dimension	Cohen’s κ	Krippendorff’s α
ISS	0.8203	0.8216
JUS	0.7036	0.7047
DIS	0.8909	0.8919

Table 6: Dimension-specific agreement between Qwen3-Plus and the human majority annotations.

expression order, or level of detail. They describe the same disputed issue, an equivalent key reasoning point, or the same principal administrative disposition.

- **Partially Equivalent:** The model-output extract and the ground-truth content partially overlap in legal meaning, but the model-output extract omits some salient elements, contains minor deviations, or expresses only part of the relevant legal information.
- **Not Equivalent:** The model-output extract and the ground-truth content concern different disputed issues, reasoning points, or administrative dispositions and therefore cannot be regarded as semantically consistent in legal meaning.

For each validation triplet, the majority label assigned by the three legal professionals is treated as the human annotation and is compared with the label produced by Qwen3-Plus.

Evaluation Metrics. We report Cohen’s κ and Krippendorff’s α separately for ISS, JUS, and DIS. Both metrics are computed between the label sequence produced by Qwen3-Plus and the corresponding sequence of human majority labels for each dimension.

Results and Error Analysis. As shown in Table 6, Qwen3-Plus achieves consistently high agreement with the human majority annotations across the three semantic matching dimensions. The highest agreement is observed for DIS, with a Cohen’s κ of 0.8909 and a Krippendorff’s α of 0.8919. For ISS, the corresponding scores are 0.8203 and 0.8216. JUS is more challenging, but Qwen3-Plus still obtains a Cohen’s κ of 0.7036 and a Krippendorff’s α of 0.7047.

We manually examine all 17 validation instances in which the Qwen3-Plus label differs from the human majority label. The main patterns of disagreement are as follows:

1. **Alternative but legally plausible reasoning in JUS.** In some validation triplets, the model-output extract and the ground-truth content follow different but legally plausible reasoning paths given the source legal text. Although these reasoning points may support similar legal conclusions, determining whether they should be classified as fully equivalent, partially equivalent, or not equivalent can be inherently ambiguous. This type of disagreement helps explain the lower agreement observed for JUS compared with ISS and DIS.
2. **Differences in disposition granularity in DIS.** A small number of disagreements arise when the model-output extract and the ground-truth content convey the same principal administrative disposition but differ in their level of granularity. For example, one may include additional orders, conditions, or implementation details, while the other captures only the principal disposition. Such differences can result in borderline judgments between *Equivalent* and *Partially Equivalent*.

Conclusion. The validation results show that Qwen3-Plus achieves substantial agreement with human majority annotations across ISS, JUS, and DIS, with particularly high agreement for ISS and DIS. The remaining disagreements are limited and largely attributable to interpretable edge cases, particularly alternative legal reasoning paths in JUS. These results suggest that Qwen3-Plus can serve as a practical substitute for human semantic matching for these three dimensions under our evaluation setting, supporting scalable and consistent evaluation in ChJGBench.

E Legal Ability Family Taxonomy

To support a more interpretable analysis beyond individual metrics, we group the 13 evaluation metrics into four legal ability families, as summarized in Table 7. This grouping allows us to examine whether model weaknesses are confined to individual tasks or extend across broader legal capabilities. The family assignments determine the point colors in Figure 2, allowing task difficulty and cross-model variability to be interpreted at the capability level. The statistical construction of the figure is described in Appendix F.

Family	Metrics
Legal Basis	Law
Issue & Qualification	Charge, ISS SUP _{Adm} , SUP _{Civ}
Judicial Reasoning	JUS, REC MIT
Disposition & Quantification	Fine, Pen_M, DIS PAY _{Adm} , PAY _{Civ}

Table 7: Legal ability family taxonomy used for task-family analysis.

F Task Difficulty and Cross-Model Variability

This appendix describes the statistical construction of Figure 2. Each task t corresponds to a particular evaluation metric for criminal, civil, or administrative cases and is assigned to a legal ability family according to Appendix E. Let μ_{tm} and σ_{tm} denote the mean and sample standard deviation over three runs, respectively, for model m on task t , as reported in Table 3. With $M = 9$, the horizontal coordinate is

$$x_t = \frac{1}{M} \sum_{m=1}^M \mu_{tm}. \quad (8)$$

The vertical coordinate is

$$y_t = \sqrt{\frac{1}{M-1} \sum_{m=1}^M (\mu_{tm} - x_t)^2}. \quad (9)$$

The dashed vertical and horizontal lines mark the medians of x_t and y_t , respectively, across the 19 tasks. Assuming independent run-to-run variation across models, the horizontal error-bar half-width is

$$u_{x,t} = \frac{1}{M} \sqrt{\sum_{m=1}^M \sigma_{tm}^2}. \quad (10)$$

The vertical error-bar half-width is

$$u_{y,t} = \sqrt{\sum_{m=1}^M \left(\frac{(\mu_{tm} - x_t)\sigma_{tm}}{(M-1)y_t} \right)^2}. \quad (11)$$

The error bars represent propagated run-to-run variability rather than confidence intervals. Half-widths below 0.005 are omitted independently on each axis.

G Benchmark Dataset Statistics

Table 8 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 includes a reference output comprising the *Reasoning* and *Judgment* sections. The complete reference output, *Reasoning* section, and *Judgment* section have average lengths of 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.

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

Table 8: Basic statistics of the benchmark dataset. Abbreviations: average prompt length (Prompt len.), average length of the *Reasoning* section (Reason len.), average length of the *Judgment* section (Judgment len.), and average number of cited legal provisions (Cited provisions).

H Generation Prompt

Under the same experimental conditions, all models receive the same system instruction, three demonstrations in the same order, target-input format, and output constraints. The generation input contains only the *Header*, *Background*, and *Findings of Fact*, while the model is required to generate only the *Reasoning* and *Judgment*.

Generation Prompt (3-shot Format)

System Message

You are completing a research benchmark task for Chinese legal judgment generation. Given the Header, Background, and Findings of Fact from a Chinese judgment, generate only its Reasoning and Judgment sections in Chinese.

The Reasoning is the public-facing judicial analysis that would appear in the written judgment. It should explain how the court legally characterizes the established facts, addresses legally material claims, defenses, charges, sentencing circumstances, or disputed issues when applicable, and applies relevant law to support the disposition. It should present the court's stated legal reasoning rather than any private or hidden reasoning process. Where appropriate, cite the relevant statutory provisions

and conclude by leading naturally to the operative ruling.

The Judgment states the court's final operative disposition . It should clearly specify the legal consequences for the relevant parties, such as conviction and sentencing, payment or performance obligations, disposition of civil claims, or the validity, annulment, modification, or required performance of administrative acts. When multiple parties, claims, charges, or obligations are involved, state their respective outcomes separately and unambiguously.

Task and Requirements

1. Analyze the legally material facts and issues and explain how the applicable law supports the disposition.
2. Address material claims, defenses, charges, sentencing issues, payment obligations, or administrative remedies only when applicable and supported by the input.
3. Do not introduce or alter parties, facts, evidence, dates, amounts, claims, or procedural events. Preserve anonymized names and identifiers exactly as provided.
4. Cite only relevant statutory provisions, using their official titles and article numbers. Do not fabricate uncertain citations.
5. Address multiple parties, defendants, claims, charges, or obligations separately and unambiguously.
6. Write in formal and coherent Chinese judicial style.
7. Output only the Reasoning and Judgment sections. Do not add task explanations, evaluation comments, confidence scores, or Markdown fences.

Demonstration 1

Input

Header:
{demo_1_header}

Background:
{demo_1_background}

Findings of Fact:
{demo_1_findings_of_fact}

Output

Reasoning:
{demo_1_reasoning}

Judgment:
{demo_1_judgment}

Demonstration 2

Input

Header:
{demo_2_header}

Background:
{demo_2_background}

Findings of Fact:
{demo_2_findings_of_fact}

Output

Reasoning:
{demo_2_reasoning}

Judgment:
{demo_2_judgment}

Demonstration 3

Input

Header:
{demo_3_header}

Background:

{demo_3_background}

Findings of Fact:
{demo_3_findings_of_fact}

Output

Reasoning:
{demo_3_reasoning}

Judgment:
{demo_3_judgment}

Target Case

Input

Header:
{target_header}

Background:
{target_background}

Findings of Fact:
{target_findings_of_fact}

Required Output

Reasoning:

Judgment:

I Dimension-Specific Extraction and Classification: Prompts and Scoring Procedures

We provide the dimension-specific extraction and classification prompts for all 13 evaluation dimensions, together with the corresponding deterministic scoring procedures. The semantic matching and scoring procedures for ISS, JUS, and DIS are presented separately in Appendix J. These prompts are used by Qwen3-Plus for controlled information extraction and classification. Braced expressions such as {text} denote runtime placeholders.

I.1 Shared Dimensions

Law: Statutory Provision Application

Task and Requirements

I will provide the Court's Reasoning section of a model-generated judgment. Extract every statutory provision explicitly cited in the input text.

Requirements:

1. Extract only provisions explicitly present in the text. Do not infer or add uncited provisions.
2. Normalize each provision to the form "[Statute Name] Article X", while retaining the official statute name in the source language.
3. When one citation contains multiple article numbers, return each combination of the statute name and an article number separately.
4. Return each unique provision once, in the order in which it first appears.
5. If no statutory provision is explicitly cited, return an empty list [].
6. Return only the required list without any explanation.

Output Format

Return a strict List[str] in the following form: ["[Statute Name] Article 234", "[Statute Name] Article 67"]

Input

```
{text}
```

Extract the results according to the requirements above:

The extracted provisions are matched against the provisions in the ground truth. Two items are considered a match when their article numbers are identical and their statute names are identical or belong to the same predefined alias group. Duplicate items are counted only once.

For each case c , let TP_c denote the number of matched provisions, FP_c the number of provisions appearing only in the extracted result, and FN_c the number appearing only in the ground truth. The case-level Law score is

$$F_{1,c}^{\text{Law}} = \frac{2TP_c}{2TP_c + FP_c + FN_c}. \quad (12)$$

If the ground truth contains at least one provision but the extracted result for the model-generated judgment is empty, the case-level score is 0.

Let K denote the number of evaluated cases. The final Law score is the mean of the case-level scores:

$$\text{Law} = \frac{1}{K} \sum_{c=1}^K F_{1,c}^{\text{Law}}. \quad (13)$$

ISS: Disputed Issues Identification

Task and Requirements

Task: I will provide the Reasoning section of a judgment. Extract a list of the disputed issues addressed in that section.

Definition: A disputed issue is an outcome-relevant factual or legal question that the court must resolve. It may concern rights and obligations between parties in a civil case, criminal liability or prosecution in a criminal case, or the legality of an administrative action in an administrative case; it need not be framed as an express disagreement between two private parties.

Requirements:

1. Extract only issues supported by the input text; do not infer or invent issues.
2. Use the finest reasonable granularity. Each item must express one indivisible, atomic disputed issue.
3. Preserve the legally material parties, acts, amounts, and conditions when they are necessary to identify an issue.
4. If no disputed issue is stated or can be directly extracted, return an empty list [].

Output Format

Return a strict List[str], for example:

```
[
  "Whether the agreement reflects the parties' true intent
  and is legally valid",
  "Whether the extension of the construction period
  constitutes a breach by the defendant",
  "Whether the plaintiff's claim for CNY 200,000 in
  liquidated damages has a factual and legal basis"
]
```

Input
{text}

Extract the results according to the requirements above:

JUS: Judicial Justification

Task and Requirements

Task: I will provide the Reasoning section of a judgment. Extract a list of the court's key judicial justifications for all material conclusions, including the treatment of claims, determinations of criminal liability and sentencing, and administrative dispositions, as applicable.

Definition: A judicial justification is the court's analysis explaining why a particular legal conclusion or disposition follows from the established facts, disputed issues, and applicable law.

Requirements:

1. Extract only reasoning points supported by the input text; do not add new facts, legal grounds, or conclusions.
2. Use the finest reasonable granularity. Each item must express one indivisible, atomic reasoning point.
3. Retain the factual premise, legal basis, and resulting conclusion when they are explicitly connected in the text.
4. Do not treat a bare fact, statutory citation, or dispositive statement as a justification unless the text uses it as part of the court's reasoning.
5. If no judicial justification can be directly extracted, return an empty list [].

Output Format

Return a strict List[str], for example:
["<reason 1>", "<reason 2>", "<reason 3>"]

Input
{text}

Extract the results according to the requirements above:

I.2 Criminal Case Dimensions

Charge: Criminal Charge Prediction

Task and Requirements

I will provide the Judgment section of a model-generated criminal judgment. Extract the criminal charge or charges of which each defendant is convicted.

Requirements:

1. Extract only criminal charges explicitly stated in the input text.
2. Use each defendant's name as a dictionary key and a list of that defendant's charge labels as the value.
3. Preserve charge labels in the source language.
4. Do not include charges stated only as allegations if the final judgment does not adopt them.
5. When a defendant is convicted of multiple charges, return each charge separately.
6. Do not include civil causes of action, administrative violations, sentencing factors, or punishment types.
7. If a defendant is identified but no charge of conviction can be extracted, return an empty list for that defendant.
8. If no defendant-specific charge information is present, return an empty dictionary {}.
9. Return only the required dictionary without any explanation.

Output Format

Return a strict Python dictionary in the following form:
{"<defendant_name>": [<charge_1>, "<charge_2>"]}

Example:

```
{"Defendant A": ["Charge 1", "Charge 2"], "Defendant B":
  []}
```

Input
{text}

Extract the results according to the requirements above:

The extracted results are matched to the ground

truth by defendant name. For each defendant in the ground truth, a charge appearing in both the extracted result and the ground truth is counted as a true positive. A charge appearing only in the extracted result is counted as a false positive, while a charge appearing only in the ground truth is counted as a false negative. Defendants appearing only in the extracted result are ignored.

For defendant i in case c , let $TP_{c,i}$, $FP_{c,i}$, and $FN_{c,i}$ denote the corresponding numbers of true positives, false positives, and false negatives. The defendant-level Charge score is

$$F_{1,c,i}^{\text{Charge}} = \frac{2TP_{c,i}}{2TP_{c,i} + FP_{c,i} + FN_{c,i}}. \quad (14)$$

If a defendant in the ground truth has no corresponding extracted result or has an empty extracted charge list, the defendant-level F_1 score is 0.

Let N_c denote the number of defendants in the ground truth for case c . The case-level Charge score is

$$S_c^{\text{Charge}} = \frac{1}{N_c} \sum_{i=1}^{N_c} F_{1,c,i}^{\text{Charge}}. \quad (15)$$

Let K denote the number of evaluated criminal cases. The final Charge score is

$$\text{Charge} = \frac{1}{K} \sum_{c=1}^K S_c^{\text{Charge}}. \quad (16)$$

Fine: Monetary Fine

Task and Requirements

I will provide the Judgment section of a criminal judgment.
Extract the monetary fine imposed on each defendant.

Definition: A monetary fine is a criminal financial punishment imposed by the court. It does not include compensation, restitution, refunds, civil payment obligations, confiscation, litigation costs, or other monetary items.

Requirements:

1. Extract only fines explicitly imposed in the input text. Do not infer an amount.
2. Use each defendant's name as a dictionary key and the fine amount as the value.
3. Normalize every amount to an integer number of Chinese yuan and omit the currency unit and all other text.
4. If the judgment expressly states a final aggregate fine for a defendant, return that aggregate and do not double-count its component amounts.
5. If a defendant is mentioned but no fine is imposed, or a fine is mentioned without a numerical amount, return None for that defendant.
6. If no defendant-specific fine information is present, return an empty dictionary {}.

Output Format

Return a strict Python dictionary in the following form:
{ "<defendant_name>": <int or None> }

Example:

{ "Defendant A": 5000, "Defendant B": None }

```
### Input
{text}
```

Extract the results according to the requirements above:

We apply the extraction prompt to the model-generated judgment to obtain structured fine predictions. The extracted results are matched to the ground truth by defendant name. Each defendant with a ground-truth fine is included in the evaluation. When a numerical predicted amount is available for the same defendant, the score is calculated using Equation 5. If the predicted value is missing, the corresponding defendant-level score is set to 0. Defendants appearing only in the prediction are ignored.

Let n_c denote the number of defendants with a ground-truth fine in case c , and let $s_{c,i}^{\text{Fine}}$ denote the score for the i -th defendant. When a numerical prediction $a_{c,i}$ is available, $s_{c,i}^{\text{Fine}} = \text{NVS}(a_{c,i}, b_{c,i})$, where $b_{c,i}$ is the ground-truth amount. When the predicted value is missing, $s_{c,i}^{\text{Fine}} = 0$. The case-level Fine score is

$$S_c^{\text{Fine}} = \frac{1}{n_c} \sum_{i=1}^{n_c} s_{c,i}^{\text{Fine}}. \quad (17)$$

Let K denote the number of cases containing at least one defendant with a ground-truth fine. The final Fine score is

$$\text{Fine} = \frac{1}{K} \sum_{c=1}^K S_c^{\text{Fine}}. \quad (18)$$

REC: Recidivism Assessment

Task and Requirements

I will provide the Reasoning section of a model-generated criminal judgment. Determine, for each defendant, whether the court identifies that defendant as a recidivist.

Requirements:

1. Return True if the text explicitly identifies the defendant as a recidivist; otherwise return False.
2. Base the result only on the input text. Do not infer recidivism from prior convictions, repeated offending, or a criminal record unless the court explicitly states the legal determination.
3. Use each defendant's name as a dictionary key and a Boolean value as the result.
4. If no defendant is identifiable in the input, return an empty dictionary {}.
5. Return only the required dictionary without any explanation.

Output Format

Return a strict Python dictionary in the following form:
{ "<defendant_name>": <True or False> }

Example:

{ "Defendant A": True, "Defendant B": False }

Input

{text}

Extract the results according to the requirements above:

The extracted results are matched to the ground truth by defendant name. Each defendant in the ground truth is treated as one classification instance, and if no corresponding prediction is available, the predicted label for that defendant is set to False. Defendants appearing only in the extracted result are ignored.

The labels from all evaluated criminal cases are pooled before scoring, and no case-level averaging is performed. Let TP, FP, FN, and TN denote the numbers of true positives, false positives, false negatives, and true negatives, respectively, with True representing the recidivist class. The F_1 score for the recidivist class is

$$F_1^{\text{True}} = \frac{2TP}{2TP + FP + FN}, \quad (19)$$

and the F_1 score for the non-recidivist class is

$$F_1^{\text{False}} = \frac{2TN}{2TN + FP + FN}. \quad (20)$$

The final REC score is the macro-average of the two class-level scores:

$$\text{REC} = \frac{F_1^{\text{True}} + F_1^{\text{False}}}{2}. \quad (21)$$

MIT: Mitigating-Circumstance Identification

Task and Requirements

I will provide the Reasoning section of a model-generated criminal judgment. Determine, for each defendant, whether the court recognizes any sentencing circumstance that supports lenient treatment.

Requirements:

1. Return True if the text explicitly recognizes at least one circumstance supporting lenient treatment for the defendant; otherwise return False.
2. Such circumstances may include voluntary surrender, confession, remorse, restitution, compensation, or obtaining the victim's forgiveness.
3. Base the result only on the input text. Do not infer a mitigating circumstance from conduct that the court does not explicitly recognize as supporting lenient treatment.
4. Use each defendant's name as a dictionary key and a Boolean value as the result.
5. If no defendant is identifiable in the input, return an empty dictionary {}.
6. Return only the required dictionary without any explanation.

Output Format

Return a strict Python dictionary in the following form: {"<defendant_name>": <True or False>}

Example:

{"Defendant A": True, "Defendant B": False}

Input

{text}

Extract the results according to the requirements above:

The extracted results are matched to the ground truth by defendant name. Each defendant in the ground truth is treated as one classification instance, and if no corresponding prediction is available, the predicted label for that defendant is set to False. Defendants appearing only in the extracted result are ignored.

The labels from all evaluated criminal cases are pooled before scoring, and no case-level averaging is performed. Let TP, FP, FN, and TN denote the numbers of true positives, false positives, false negatives, and true negatives, respectively, with True indicating that at least one mitigating circumstance is recognized. The F_1 score for the positive class is

$$F_1^{\text{True}} = \frac{2TP}{2TP + FP + FN}, \quad (22)$$

and the F_1 score for the negative class is

$$F_1^{\text{False}} = \frac{2TN}{2TN + FP + FN}. \quad (23)$$

The final MIT score is the macro-average of the two class-level scores:

$$\text{MIT} = \frac{F_1^{\text{True}} + F_1^{\text{False}}}{2}. \quad (24)$$

Pen_M: Penalty Duration Prediction

Task and Requirements

I will provide the sentencing portion of a criminal judgment. Extract the duration of the principal punishment imposed on each defendant and normalize it to months.

Requirements:

1. Extract the final duration of a duration-based principal punishment, including fixed-term imprisonment, criminal detention, or public surveillance.
2. Convert years and months to a total number of months. For example, three years is 36 months, and one year and six months is 18 months.
3. If both individual sentences and an expressly imposed combined sentence are stated, extract the duration of the combined principal punishment.
4. If the punishment is life imprisonment, return the string "LIFE_IMPRISONMENT".
5. If the punishment is the death penalty, including a death sentence with a reprieve, return the string "DEATH_PENALTY".
6. If the punishment is exemption from criminal punishment, or if no duration-based principal punishment is stated, return None.
7. If a suspended sentence is imposed for a duration-based principal punishment, extract the duration of the principal punishment and ignore the probation period.

Output Format

Return a strict Python dictionary using each defendant's name as the key: {"<defendant_name>": <integer_months or "LIFE_IMPRISONMENT" or "DEATH_PENALTY" or None>}

Example input:

Defendant Gao is sentenced to one year of fixed-term imprisonment, suspended for two years; Defendant Li is sentenced to life imprisonment; Defendant Wang is sentenced to the death penalty with a two-year reprieve.

Example output:
{"Defendant Gao": 12, "Defendant Li": "LIFE_IMPRISONMENT",
"Defendant Wang": "DEATH_PENALTY"}

Input
{text}

Extract the results according to the requirements above:

We apply the extraction prompt to the model-generated judgment to obtain structured punishment predictions. The extracted results are matched to the ground truth by defendant name. If the ground truth provides a numerical duration or a categorical punishment for a defendant but no corresponding punishment is extracted from the model-generated judgment, the defendant receives a score of 0 according to Equation 6.

Let N_c denote the number of defendants for whom the ground truth provides such a punishment in case c , and let $s_{c,i}^{\text{Pen_M}}$ denote the score assigned to the i -th defendant according to Equation 6. The case-level Pen_M score is

$$S_c^{\text{Pen_M}} = \frac{1}{N_c} \sum_{i=1}^{N_c} s_{c,i}^{\text{Pen_M}}, \quad (25)$$

where $N_c > 0$. Let K denote the number of cases with $N_c > 0$. The final Pen_M score is

$$\text{Pen_M} = \frac{1}{K} \sum_{c=1}^K S_c^{\text{Pen_M}}. \quad (26)$$

I.3 Civil Case Dimensions

SUP_{Civ}: Civil Claim Support

Task and Requirements

I will provide the Reasoning and Judgment sections of a civil judgment. Determine the overall disposition of the plaintiff's claims.

Requirements:

1. Classify the overall disposition as exactly one of: "upheld", "partially upheld", or "rejected".
2. Return "upheld" when all claims decided by the court are fully upheld.
3. Return "rejected" when none of the claims decided by the court are upheld.
4. Return "partially upheld" when the plaintiff receives some but not complete support, including when a claim is only partially upheld or when multiple claims receive different outcomes.
5. Determine the outcome solely from the input text.
6. If the Reasoning and Judgment sections differ, follow the final operative Judgment section.
7. Return only the required dictionary without any explanation.

Output Format

Return a strict Python dictionary in the following form:
{"outcome": "<upheld or partially upheld or rejected>"}

Input
{text}

Determine the overall disposition according to the requirements above:

For each case, the predicted label is compared with the ground-truth label. For each outcome label, let P and R denote precision and recall, respectively. Its F_1 score is

$$F_1 = \frac{2PR}{P + R}. \quad (27)$$

In our evaluations, $P + R > 0$ for all three outcome labels. Let F_1^U , F_1^{PU} , and F_1^R denote the F_1 scores for upheld, partially upheld, and rejected, respectively. The final SUP_{Civ} score is their un-weighted mean:

$$\text{SUP}_{\text{Civ}} = \frac{F_1^U + F_1^{PU} + F_1^R}{3}. \quad (28)$$

PAY_{Civ}: Civil Payment Obligation

Task and Requirements

I will provide the Judgment section of a civil judgment. Extract the total amount of civil payment obligations imposed on each obligor.

Requirements:

1. Extract only monetary payment obligations explicitly imposed in the final Judgment section.
2. Include compensation, repayment, damages, restitution, and other civil monetary liabilities.
3. Do not extract criminal fines, litigation costs, or non-monetary performance obligations.
4. Use the name of the person or entity bearing the payment obligation as the dictionary key.
5. If a final total is explicitly stated for an obligor, return that total. Otherwise, sum the explicitly stated payment amounts imposed on that obligor.
6. Normalize each amount to an integer number of Chinese yuan and omit the currency unit and all other text.
7. If an obligor is identified but no numerical amount can be extracted, return None for that obligor.
8. If no civil payment obligation is imposed, return an empty dictionary {}.
9. Return only the required dictionary without any explanation.

Output Format

Return a strict Python dictionary in the following form:
{"<civil_obligor_name_or_entity>": <int or None>}

Example:

{"Company A": 5000, "Party B": None}

Input

{text}

Extract the results according to the requirements above:

The extracted amounts are matched to the ground truth by obligor name. For each obligor with both a predicted amount and a ground-truth amount, the NVS score is calculated using Equation 5. If an obligor has a ground-truth amount but the predicted value is missing, the corresponding NVS score is set to 0. Obligor appearing only in the prediction are ignored. PAY is evaluated only on cases containing at least one reference payment obligation. Let n_c denote the number of obligors with a payment obligation in the reference judgment for

civil case c , and let $s_{c,i}$ denote the score of the i -th obligor. The case-level score is

$$S_c^{\text{PAY}_{\text{Civ}}} = \frac{1}{n_c} \sum_{i=1}^{n_c} s_{c,i}. \quad (29)$$

Let K denote the number of evaluated civil cases. The final PAY_{Civ} score is

$$\text{PAY}_{\text{Civ}} = \frac{1}{K} \sum_{c=1}^K S_c^{\text{PAY}_{\text{Civ}}}. \quad (30)$$

I.4 Administrative Case Dimensions

SUP_{Adm}: Administrative Claim Support

Task and Requirements

I will provide the Reasoning and Judgment sections of an administrative judgment. Determine the overall disposition of the plaintiff's claims.

Requirements:

1. Classify the overall disposition as exactly one of: "upheld", "partially upheld", or "rejected".
2. Return "upheld" when all claims decided by the court are fully upheld.
3. Return "rejected" when none of the claims decided by the court are upheld.
4. Return "partially upheld" when the plaintiff receives some but not complete support, including when a claim is only partially upheld or when multiple claims receive different outcomes.
5. Classify the outcome of the plaintiff's claims, not whether the challenged administrative act remains in effect.
6. Determine the outcome solely from the input text.
7. If the Reasoning and Judgment sections differ, follow the final operative Judgment section.
8. Return only the required dictionary without any explanation.

Output Format

Return a strict Python dictionary in the following form: {"outcome": "<upheld or partially upheld or rejected>"}

Input
{text}

Determine the overall disposition according to the requirements above:

2. Include administrative compensation, restitution, refunds, and other monetary obligations ordered by the court.
3. Do not extract a monetary fine merely because it appears in the challenged administrative act.
4. Do not extract criminal fines, litigation costs, or non-monetary administrative dispositions.
5. Use the name of the person, entity, or administrative authority bearing the payment obligation as the dictionary key.
6. If a final total is explicitly stated for an obligor, return that total. Otherwise, sum the explicitly stated payment amounts imposed on that obligor.
7. Normalize each amount to an integer number of Chinese yuan and omit the currency unit and all other text.
8. If an obligor is identified but no numerical amount can be extracted, return None for that obligor.
9. If no administrative payment obligation is imposed, return an empty dictionary {}.
10. Return only the required dictionary without any explanation.

Output Format

Return a strict Python dictionary in the following form: {"<administrative_obligor_name_or_entity>": <int or None>}

Example:

{"Administrative Authority A": 5000, "Entity B": None}

Input
{text}

Extract the results according to the requirements above:

The extracted amounts are matched to the ground truth by obligor name. For each obligor with both a predicted amount and a ground-truth amount, the NVS score is calculated using Equation 5. If an obligor has a ground-truth amount but the predicted value is missing, the corresponding NVS score is set to 0. Obligor appearing only in the prediction are ignored. PAY is evaluated only on cases containing at least one reference payment obligation. Let n_c denote the number of obligors with a payment obligation in the reference judgment for administrative case c , and let $s_{c,i}$ denote the score of the i -th obligor. The case-level score is

$$S_c^{\text{PAY}_{\text{Adm}}} = \frac{1}{n_c} \sum_{i=1}^{n_c} s_{c,i}. \quad (32)$$

Let K denote the number of evaluated administrative cases. The final PAY_{Adm} score is

$$\text{PAY}_{\text{Adm}} = \frac{1}{K} \sum_{c=1}^K S_c^{\text{PAY}_{\text{Adm}}}. \quad (33)$$

For each case, the predicted label is compared with the ground-truth label. The label-level F_1 scores are calculated using Equation 27. Let F_1^U , F_1^{PU} , and F_1^R denote the F_1 scores for upheld, partially upheld, and rejected, respectively. The final SUP_{Adm} score is their unweighted mean:

$$\text{SUP}_{\text{Adm}} = \frac{F_1^U + F_1^{PU} + F_1^R}{3}. \quad (31)$$

PAY_{Adm}: Administrative Payment Obligation

Task and Requirements

I will provide the Judgment section of an administrative judgment. Extract the total amount of administrative payment obligations imposed on each obligor.

Requirements:

1. Extract only monetary payment obligations explicitly imposed in the final Judgment section.

DIS: Administrative Disposition Prediction

Task and Requirements

Task: I will provide the Judgment section of an administrative judgment. Extract a list of the core non-monetary administrative dispositions made by the court concerning the challenged administrative act or an administrative authority's statutory duties.

Definition: An administrative disposition is a judicial determination concerning the legal effect of a challenged administrative act or an administrative

authority's ensuing non-monetary duties. It includes, but is not limited to, annulment or partial annulment of an administrative act, upholding an administrative act, declaring an administrative act unlawful or invalid, modifying an administrative act, ordering an administrative authority to perform a statutory duty, ordering an authority to issue a new administrative act, and other non-monetary dispositions with equivalent legal effect.

Requirements:

1. Extract only dispositions explicitly stated in the input text. Do not infer, supplement, or rewrite them based on the facts of the case.
2. Each list item must contain one indivisible, atomic disposition. If the text concerns multiple administrative acts, authorities, or dispositions, list them separately.
3. Retain enough information to identify each disposition, including the dispositive action, its object, and any expressly stated scope or condition, such as "in part", "within the statutory time limit", or "issue a new administrative act". Do not return context-free labels such as "annulled" or "upheld" alone.
4. If dismissal of the plaintiff's claims is the only explicit non-monetary outcome, extract that outcome as written; do not rewrite it as upholding the administrative act.
5. Do not extract administrative compensation, restitution, refunds, monetary fines, or other payment obligations under this dimension. A determination that annuls, upholds, or declares unlawful an administrative penalty decision involving a fine is nevertheless a non-monetary disposition and should be extracted.
6. Do not extract litigation-cost allocations, appeal deadlines, or other procedural information.
7. If no explicit non-monetary administrative disposition is present, return an empty list [].

Output Format

Return a strict List[str], for example:

```
[
  "Annul Administrative Penalty Decision No. [X] issued by
  the defendant authority",
  "Order the defendant authority to issue a new
  administrative act within the statutory time limit"
]
```

Input

{text}

Extract the results according to the requirements above:

J Semantic Matching Prompts and Deterministic Scoring

For each dimension $d \in \{\text{ISS}, \text{JUS}, \text{DIS}\}$, Qwen3-Plus receives the source case context and lists of model-output extracts and ground-truth items. In the templates below, *source_case_context* contains only the *Header*, *Background*, and *Findings of Fact* provided to the judgment-generation model. Following the dimension-specific rubric, Qwen3-Plus assigns *Equivalent*, *Partially Equivalent*, or *Not Equivalent* to every possible pair consisting of one model-output extract and one ground-truth item. These tags are mapped to fixed weights:

$$w(t) = \begin{cases} 1, & t = \text{Equivalent}, \\ 0.5, & t = \text{Partially Equivalent}, \\ 0, & t = \text{Not Equivalent}. \end{cases} \quad (34)$$

For each case c , let m_c and n_c denote the numbers of ground-truth content items and model-output extracts, respectively. Let T_i denote the i -th ground-truth content item and M_j denote the j -th model-output extract. Let $t_{ij}^{(c)}$ denote the semantic tag assigned to the pair consisting of T_i and M_j , and let $w(t_{ij}^{(c)})$ be its corresponding weight defined in Eq. 34. We then construct a complete bipartite graph between the m_c ground-truth content items and the n_c model-output extracts, where the edge between T_i and M_j has weight $w(t_{ij}^{(c)})$.

Let $\mathcal{M}_c$ denote the set of all valid one-to-one matchings between the ground-truth content items and the model-output extracts for case c , such that each item can participate in at most one matched pair. We define $W_{d,c}$ as the maximum total semantic weight achievable under such a matching:

$$W_{d,c} = \max_{M \in \mathcal{M}_c} \sum_{(i,j) \in M} w(t_{ij}^{(c)}). \quad (35)$$

The case-level soft semantic F_1 score is then defined as

$$S_{d,c} = \begin{cases} 1, & m_c = n_c = 0, \\ 0, & \text{exactly one of } m_c, n_c \text{ is } 0, \\ \frac{2W_{d,c}}{m_c + n_c}, & \text{otherwise.} \end{cases} \quad (36)$$

Let N_d denote the number of applicable cases for dimension d . The final score for dimension d is computed as the arithmetic mean of the case-level soft semantic F_1 scores:

$$\text{Score}_d = \frac{1}{N_d} \sum_{c=1}^{N_d} S_{d,c}. \quad (37)$$

This case-level averaging gives equal weight to each applicable case, regardless of the numbers of ground-truth content items and model-output extracts in that case. The scoring method gives partial credit for partial semantic equivalence while penalizing both omitted ground-truth content and unsupported model-output extracts.

J.1 ISS: Disputed Issues Identification

ISS: Pairwise Semantic Tagging

Task

Compare every ground-truth disputed issue with every model-output disputed issue. Each input item has a unique identifier. Assign exactly one semantic-equivalence tag to every possible pair consisting of one ground-truth disputed issue and one model-output disputed issue.

```

### Definition
A disputed issue is an outcome-relevant factual or legal
question that the court must resolve. It may concern
rights and obligations between parties in a civil
case, a criminal accusation or defense in a criminal
case, or the legality of an administrative act or
omission in an administrative case. An express
disagreement between two private parties is not
required.

### Tags
- "Equivalent": The two items express the same outcome-
relevant factual or legal question. Differences in
wording, order, abstraction, or non-material detail
do not change the question to be adjudicated.
- "Partially Equivalent": The two items substantially
concern the same disputed question, but one is
broader or narrower, or omits or changes a legally
salient party, act, object, amount, condition, time
period, or requested remedy.
- "Not Equivalent": The two items concern different factual
or legal questions, contradict one another, or share
only general case background without addressing the
same adjudicative issue.

### Rules
1. Judge each pair independently at the atomic-issue level.
2. Compare legal meaning, not lexical overlap alone.
3. Do not tag two issues as Equivalent merely because they
arise from the same case or may lead to the same
outcome.
4. Use the source context only to resolve ambiguous parties
, acts, objects, or references. Do not create an
issue absent from either extracted item.
5. Output one comparison for every Cartesian-product pair.
Do not omit or duplicate any pair.
6. Preserve the supplied item identifiers exactly.
7. If either item list is empty, return {"comparisons":
[]}.
8. Return only the specified JSON object, with no
explanation or markdown fence.

### Output Schema
{
  "comparisons": [
    {
      "ground_truth_id": "T1",
      "model_output_id": "M1",
      "tag": "Equivalent | Partially Equivalent | Not
Equivalent"
    }
  ]
}

### Source Case Context
{source_case_context}

### Ground-Truth Disputed Issues
{ground_truth_issues}

### Model-Output Disputed Issues
{model_output_issues}

### Output

```

```

proposition, claim, defense, charge, or disposition.

### Tags
- "Equivalent": The two items express the same material
reasoning proposition. Their factual premise, legal
standard or basis, inferential relationship, and
resulting conclusion are legally compatible and
substantively the same. Differences in wording, order
, or non-material detail are permitted.
- "Partially Equivalent": The two items share a material
premise, inferential basis, or legal conclusion, but
one omits a salient fact, legal ground, analytical
step, limitation, or conclusion, or contains a minor
legally relevant deviation.
- "Not Equivalent": The two items rely on materially
different or incompatible facts, legal grounds,
reasoning paths, or conclusions. A shared fact,
citation, or final outcome alone is insufficient.

### Rules
1. Judge each pair independently at the atomic-reason level
.
2. Assess legal meaning and the fact-rule-conclusion
relationship, not prose quality or lexical overlap.
3. Do not treat a bare shared fact, statutory citation, or
disposition as sufficient for equivalence.
4. Do not tag an alternative but materially different
rationale as Equivalent merely because it is
independently plausible or reaches the same result.
5. Use the source context only to resolve ambiguity and
case-specific references. Do not add reasoning absent
from either extracted item.
6. Output one comparison for every Cartesian-product pair.
Do not omit or duplicate any pair.
7. Preserve the supplied item identifiers exactly.
8. If either item list is empty, return {"comparisons":
[]}.
9. Return only the specified JSON object, with no
explanation or markdown fence.

### Output Schema
{
  "comparisons": [
    {
      "ground_truth_id": "T1",
      "model_output_id": "M1",
      "tag": "Equivalent | Partially Equivalent | Not
Equivalent"
    }
  ]
}

### Source Case Context
{source_case_context}

### Ground-Truth Judicial-Justification Points
{ground_truth_justification_points}

### Model-Output Judicial-Justification Points
{model_output_justification_points}

### Output

```

J.2 JUS: Judicial Justification

JUS: Pairwise Semantic Tagging

```

### Task
Compare every ground-truth judicial-justification point
with every model-output judicial-justification point.
Each input item has a unique identifier. Assign
exactly one semantic-equivalence tag to every
possible pair consisting of one ground-truth judicial
-justification point and one model-output judicial-
justification point.

### Definition
A judicial-justification point is an atomic proposition in
which the court connects a legally material fact,
evidentiary assessment, or legal rule to an
intermediate or final adjudicative conclusion. It
explains why the court accepts or rejects a disputed

```

J.3 DIS: Administrative Disposition Prediction

DIS: Pairwise Semantic Tagging

```

### Task
Compare every ground-truth administrative disposition with
every model-output administrative disposition. Each
input item has a unique identifier. Assign exactly
one semantic-equivalence tag to every possible pair
consisting of one ground-truth administrative
disposition and one model-output administrative
disposition.

### Definition
An administrative disposition is a court's non-monetary
determination concerning the legal effect of a
challenged administrative act or an administrative
authority's ensuing statutory duty. Examples include
annulment or partial annulment, upholding an

```

administrative act, declaring an act unlawful or invalid, modifying an act, ordering performance of a statutory duty, and ordering an authority to issue a new administrative act.

Tags

- "Equivalent": The two dispositions have the same principal legal effect on the same administrative act, omission, authority, or statutory duty. Differences in wording, order, or non-material implementation detail are permitted.
- "Partially Equivalent": The dispositions have substantially overlapping legal effects, but one omits or changes a legally salient object, scope, condition, time limit, partial-versus-complete effect, or consequential order.
- "Not Equivalent": The dispositions concern different administrative acts or duties, contradict one another, or produce materially different legal effects, such as annulment versus upholding, performance versus dismissal, or unlawfulness versus validity.

Rules

1. Judge each pair independently at the atomic-disposition level.
2. Compare the operative legal effect, its object, and any expressly stated scope or condition.
3. Do not treat dismissal of a plaintiff's claims as automatically equivalent to upholding the challenged administrative act.
4. Exclude monetary compensation, restitution, refunds, payment obligations, litigation costs, and appeal instructions. A ruling on the validity or legal effect of an administrative penalty decision involving a fine remains a non-monetary disposition.
5. Use the source context only to resolve ambiguous acts, authorities, duties, or references. Do not create a disposition absent from either extracted item.
6. Output one comparison for every Cartesian-product pair. Do not omit or duplicate any pair.
7. Preserve the supplied item identifiers exactly.
8. If either item list is empty, return {"comparisons": []}.
9. Return only the specified JSON object, with no explanation or markdown fence.

Output Schema

```
{
  "comparisons": [
    {
      "ground_truth_id": "T1",
      "model_output_id": "M1",
      "tag": "Equivalent | Partially Equivalent | Not Equivalent"
    }
  ]
}
```

Source Case Context

```
{source_case_context}
```

Ground-Truth Administrative Dispositions

```
{ground_truth_dispositions}
```

Model-Output Administrative Dispositions

```
{model_output_dispositions}
```

Output

they primarily measure lexical overlap or semantic similarity, while LLM-as-a-judge provides only a relatively coarse holistic assessment. In contrast, our method evaluates key legal dimensions separately and produces a much lower average score, better reflecting the substantial legal inconsistency between the two judgments.

K Case Study

As shown in Figure 3, this case study illustrates how 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 amount involved. ROUGE-L and BERTScore assign relatively high scores because

Ground Truth		The Court finds that the defendant, Luo [REDACTED], for the purpose of illegal possession, used covert means to enter a residence and <u>steal</u> 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 <u>crime of theft</u> Accordingly, pursuant to Article <u>264</u> , Paragraph 3 of Article <u>67</u> , 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 <u>Handling Criminal Cases</u> of Theft ...
Candidate Answer		The Court finds that the defendant, Luo [REDACTED], for the purpose of illegal possession, used such means as concealing facts and fabricating reasons to <u>defraud</u> others of property valued at RMB 1 million. The amount involved was especially huge, and his conduct constituted the <u>crime of fraud</u> Accordingly, pursuant to Article <u>266</u> and Article <u>67</u> 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 <u>Involving Home Entry</u> ...
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-a-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 3: Case study comparing different evaluation metrics on a legally inconsistent judgment pair.