JurisBench: A Deep Benchmark for Assessing Large Language Models in Professional Legal Practice

Ziang Chen[1,2,*], Guannan Li[1,3,*], Fanlin Ji[1,3], Yipeng Kang[1], Jiaqi Li[1], Muhan Zhang[4], Yangtao Zhang[3], Li TianJiao[3], Jiannan Wang[3], Song-Chun Zhu[1,2,4], Bin Ling[3,*,✉],

[1]State Key Laboratory of General Artificial Intelligence, BIGAI,
[2]Department of Automation, Tsinghua University,
[3]Law School, Peking University, [4]Institute for Artificial Intelligence, Peking University

Correspondence: chenziang@bigai.ai, 2201111066@stu.pku.edu.cn, lingbin@pku.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated strong cross-domain capabilities, yet their competence in specialized professional tasks remains underexamined. Existing legal benchmarks evaluate isolated tasks or exam-style questions, failing to capture the *procedural interdependencies* and *adjudicative rigor* inherent in professional practice. To bridge this gap, we construct **JurisBench**, a vertical, depth-oriented, domain-specific benchmark designed to evaluate LLMs across key stages of Chinese civil litigation. JurisBench introduces a **Linear Depth Simulation** track that mirrors the cognitive workflow of professional judges through four sequential, dependency-aware phases: *Cause of Action* prediction, *Focus of Disputes* identification, *Rationale of the Judgment* generation, and *Result of the Judgment* determination. Results reveal an "illusion of competence": state-of-the-art models exhibit marked performance degradation in end-to-end pipelines due to cascading error propagation. We identify precise statutory grounding as a persistent bottleneck, highlighting a critical gap between fluent linguistic output and judicial reliability. JurisBench shifts evaluation from isolated legal knowledge to workflow-level task execution, providing a diagnostic framework for legal AI and a template for benchmark design in specialized domains.

1 Introduction

While Large Language Models (LLMs) exhibit broad knowledge spanning diverse domains, they lack deep understanding of specialized tasks inherent to professional practice. Legal reasoning exemplifies this challenge.

The rapid advancements of LLMs do demonstrate their transformative potential in the legal field (Surden, 2018; Walters and Novak, 2021; Lee et al., 2023), and recent studies report strong performance on legal question answering and knowledge-oriented tasks (Siino et al., 2025), but such results do not necessarily imply reliable performance in realistic judicial scenarios, where decisions are interdependent, procedural constraints are strict, and errors may propagate across multiple stages. Robust, practice-aligned evaluation benchmarks are therefore essential for assessing and understanding the actual capabilities and limitations of LLMs in real-world case-processing settings.

Meanwhile, the escalating volume of civil and commercial caseloads in China has placed immense pressure on judicial systems, amplifying the demand for reliable automated tools (see Appendix A). However, the complexity of real-world adjudication means that superficial legal competence—often measured by accuracy on isolated or exam-style tasks (Guha et al., 2023; Fei et al., 2023; Dai et al., 2023; Li et al., 2024b; Fan et al., 2025)—is insufficient for end-to-end judicial workflows. This mismatch between *surface-level knowledge* and *procedural reasoning* underscores the need for benchmarks that assess legal reliability across the full lifecycle of a judicial process. Existing research, while providing useful evidence of linguistic competence, often focuses on "breadth-first" task coverage at the expense of "professional depth," offering limited insight into the procedural coherence required for case-level adjudication.

To address these limitations, this paper introduces **JurisBench**[1], a vertical, depth-oriented benchmark specifically tailored to the structured judicial workflow of the Chinese legal system. JurisBench prioritizes "professional depth" by simulating the cognitive workflow of judges (Ashley, 2017) through a four-phase **Linear Depth Simulation** pipeline: *Cause of Action* prediction, *Focus of Disputes* identification, *Rationale of the Judgment*

∗ Equal contributions.
✉ Corresponding author.

[1]https://github.com/cza0927/JurisBench

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 35994–36018
July 2-7, 2026

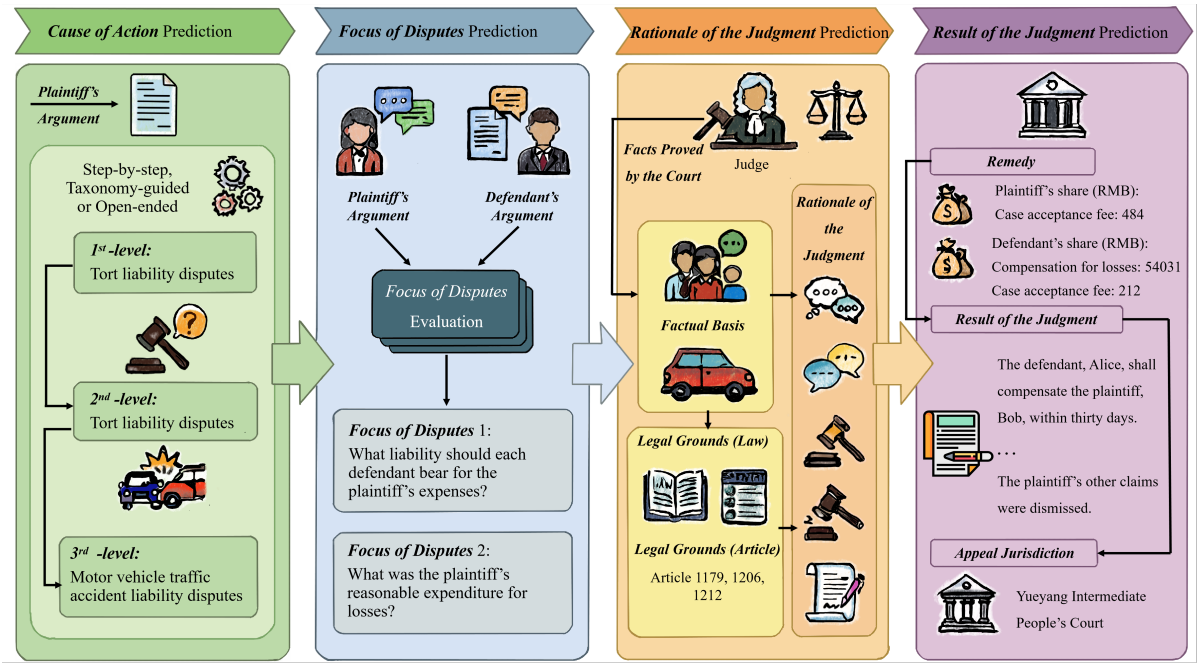


Figure 1: **An Overview of JurisBench Pipeline.** The figure illustrates the *Linear Depth Simulation* track, which models a structured judicial reasoning process for each case. The pipeline follows a decision chain: (1) Predict the *Cause of Action* based on the *Plaintiff's Argument*; (2) Identify the *Focus of Disputes* by incorporating the *Defendant's Argument*; (3) Derive the *Rationale of the Judgment* grounded in *Factual Basis* and *Legal Grounds*; and (4) Determine the *Result of the Judgment*. The logical justifications for this sequential workflow design, specifically the procedural dependencies where each phase serves as a prerequisite for the next, are elaborated in Appendix C.

generation, and *Result of the Judgment* determination (detailed in Section 3). By focusing on these interdependent tasks, JurisBench enables the systematic analysis of **reasoning stability** and **error propagation**—critical dimensions that remain invisible in traditional QA-style evaluations in the field (Zheng et al., 2023; Prince Tritto and Ponce, 2025).

Evaluating deep professional understanding differs fundamentally from testing broad legal knowledge. This objective requires focusing on specific case types with varying levels of difficulty. While JurisBench maintains broad coverage at the categorization level, its deep evaluation track focuses on *motor vehicle traffic accident liability disputes*—a prototypical form of civil litigation involving multi-party participation, factual causality analysis, and statutory grounding. This unified domain serves as a diagnostic testbed for probing professional-level reasoning depth that broader but shallower benchmarks are not designed to capture.

In summary, our contributions are as follows:

- We introduce **JurisBench**, the first deep, vertical, workflow-aligned benchmark that simulates the full lifecycle of Chinese civil litigation to evaluate the practical performance of LLMs in professional legal practice.
- We propose the **Linear Depth Simulation** track to analyze the stability of reasoning chains and the impact of cascading error propagation across interdependent judicial stages.
- Through extensive experiments on 10+ state-of-the-art LLMs, we reveal a significant gap between surface-level legal knowledge and deep judicial reasoning capability, identifying accurate comprehension of statutory grounding as a persistent bottleneck.

2 Related Work

2.1 Legal Large Language Models

LLMs have attracted substantial attention for their strong performance across diverse industries, with some claims suggesting that they can match, or may eventually replace, domain experts (Bubeck et al., 2023; Comanici et al., 2025). Such claims have encouraged both researchers and startups to invest heavily in the training and development of domain-specific LLMs, among which legal LLMs,

particularly Chinese legal LLMs, have received considerable attention.

Legal LLMs have evolved from pre-trained models such as LegalBERT (Chalkidis et al., 2020) and LexLM (Chalkidis et al., 2023) to generative systems primarily built on GPT (Walters and Novak, 2021; Lee et al., 2023; Surden, 2018) and LLaMA (Touvron et al., 2023; Kassianik et al., 2025; Hossain et al., 2025; Prince Tritto and Ponce, 2025). Early Chinese legal LLMs include ChatLaw (Cui et al., 2023), Lawyer LLaMA (Huang et al., 2023), and Lawyer GPT (Yao et al., 2024).

In practice, the construction of contemporary legal LLMs typically combines multiple adaptation strategies: **(i) domain-adaptive pretraining (DAPT)** (Chalkidis et al., 2023), **(ii) supervised fine-tuning or instruction tuning** for legal QA and case analysis (Hendrycks et al., 2021; Shen et al., 2022; Niklaus et al., 2025), and **(iii) preference-based alignment** (e.g., RLHF) (Ouyang et al., 2022; Prince Tritto and Ponce, 2025). Large-scale curated legal corpora, such as MultiLegalPile (Niklaus et al., 2024), and domain-specialized foundation models, such as SaulLM (Colombo et al., 2024), have further expanded coverage of the legal domain at scale. To improve factual reliability, retrieval-augmented generation (RAG) is widely adopted to mitigate incomplete knowledge (Lewis et al., 2020), and legal-specific pipelines further incorporate citation grounding and verification mechanisms (Qian et al., 2025; Zheng et al., 2025; Zhang et al., 2024).

However, even with such external knowledge support, a recurring reliability challenge remains in practice-aligned workflows: error propagation across multi-stage reasoning chains (Surden, 2018; Prince Tritto and Ponce, 2025). Even when authoritative materials are successfully retrieved, the application of legal rules may remain inconsistent across procedurally coupled stages (Doyle and Tucker, 2025).

2.2 Benchmarks: From General-purpose to Legal-domain

General-purpose benchmarks, such as MMLU (Hendrycks et al., 2020) and AGIEval (Zhong et al., 2024), standardize the evaluation of LLMs through closed-form tasks and outcome-oriented metrics. While effective for broad comparisons, they abstract away procedural structure and dependencies, which limits their usefulness in domains with strong structural constraints.

Existing legal benchmarks generally follow four paradigms: **(i) exam-style multi-task benchmarks**, such as LexGLUE (Chalkidis et al., 2022), LawBench (Fei et al., 2023), LEXTREME (Niklaus et al., 2023), and LEXam (Fan et al., 2025), which assess broad legal reasoning ability through standardized or exam-oriented tasks; **(ii) typology-driven or verifiability-focused benchmarks**, including LegalBench (Guha et al., 2023) and CitaLaw (Zhang et al., 2024); **(iii) depth-oriented evaluations of judicial reasoning**, such as One Law, Many Languages (Rasiah et al., 2024), which analyze reasoning structure in judicial-support settings; and **(iv) interactive or agent-based benchmarks**, such as LegalAgentBench (Li et al., 2024a), which simulate tool-augmented workflows. Despite their diversity, these benchmarks predominantly treat tasks as independent units and inherit an "exam-style" assumption that legal capability can be approximated by the aggregation of isolated outcomes.

As a result, these benchmarks underrepresent the procedural interdependence of real-world adjudication, which makes cross-stage inconsistencies difficult to identify and analyze (Surden, 2018; Prince Tritto and Ponce, 2025; Posner and Saran, 2025). This limitation motivates workflow-aligned benchmarks that evaluate procedural depth and error propagation, such as **JurisBench** proposed in this work.

3 JurisBench

This section details the systematic construction of JurisBench, including data curation, the mapping of subtasks to target judicial capabilities, complexity stratification, and evaluation metrics, as summarized in Figure 2.

3.1 Overview

JurisBench operationalizes the end-to-end cognitive process of Chinese civil litigation, mapping case filing to final judgment by using authentic judicial documents (Supreme People's Court of the PRC). To ensure ethical compliance, all records were anonymized through placeholder-based de-identification. Because missing fields are common in raw judicial records, we applied systematic filtering to ensure structural integrity. As a result, only 5.23% of the original corpus satisfied the inclusion criteria (see Appendix B). A comprehensive seman-

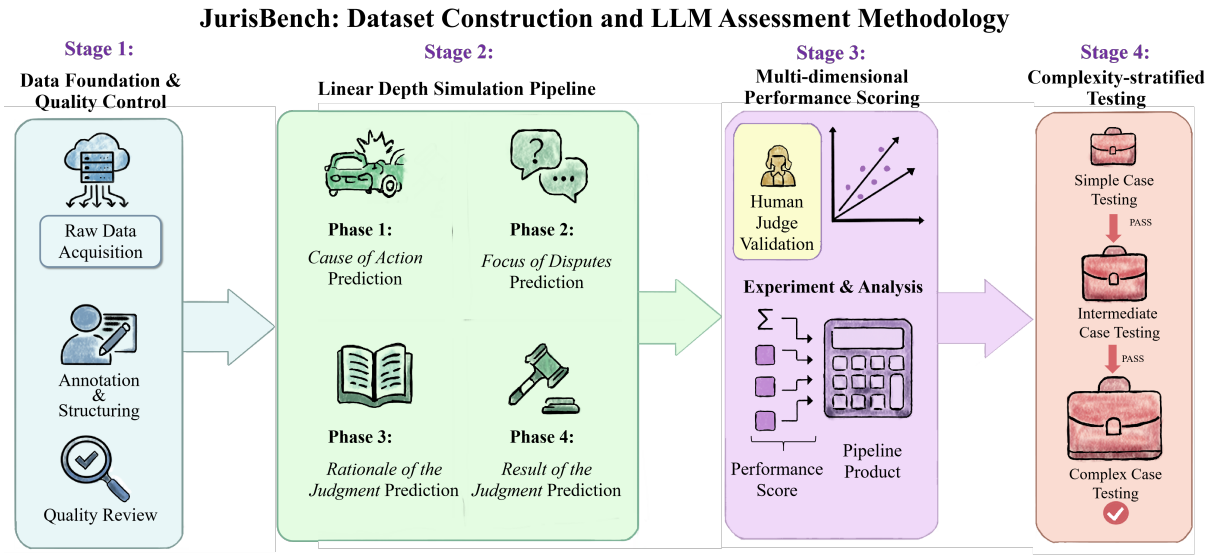


Figure 2: **Overall Methodology.** The framework proceeds from a high-quality data foundation (Stage 1) to a core Linear Depth Simulation that mirrors four judicial phases (Stage 2). Performance is then quantified through a multidimensional scoring system validated by human judges (Stage 3), followed by progressive evaluation across three complexity-stratified tiers (Stage 4).

tic analysis and t-SNE visualization, provided in Appendix D, confirm the high diversity and task variation of the JurisBench dataset across its stratified complexity levels. The benchmark adopts a *dual construction strategy* to evaluate models along two complementary dimensions (see Table 1).

Table 1: Dataset statistics of JurisBench across breadth and depth dimensions.

Evaluation Track	Complexity / Phase	Number of Cases
Depth: Pipeline Evaluation (MVTAL disputes)	Simple	394
	Intermediate	394
	Complex	212
Breadth: Taxonomy Alignment (General 282 *Causes of Action*)	Phase 1-a	2,820
	Phase 1-b	

Breadth Evaluation of *Cause of Action*. At the breadth level, the benchmark evaluates the systematic capability of models to navigate the broader landscape of Chinese civil law. This evaluation set contains 2,820 cases spanning 282 distinct *Causes of Action*. To examine the effect of informational support, this track is divided into two settings: **Phase 1-a**, in which a hierarchical candidate list from a legal database is provided in the context to simulate a restricted-selection scenario, and **Phase 1-b**, an open-ended prediction task that requires purely parametric recall. These settings assess whether models can accurately align factual narratives with formal legal ontologies.

Depth Evaluation via a Reasoning Pipeline. At the depth level, JurisBench focuses on a representative and structurally rich domain: Motor Vehicle Traffic Accident Liability (MVTAL) disputes. Evaluation is organized into four phases (see Table 2 for details), forming a dependency-aware pipeline that reflects how intermediate judicial decisions condition downstream reasoning. In parallel, standalone evaluation tracks are retained to assess independent legal knowledge. The rationale for this sequential workflow design, particularly the procedural dependencies under which each phase serves as a prerequisite for the next, is detailed in Appendix C.

3.2 Task Taxonomy and Capability Mapping

To systematically analyze the cognitive demands placed on LLMs, we map the phases of JurisBench onto five distinct capability dimensions, as detailed in Table 2. This taxonomy enables a granular diagnosis of legal reasoning. **Semantic Taxonomy Alignment** (Phase 1) assesses the ability to map informal narratives onto formal legal ontologies. **Adversarial Factual Reconstruction** (Phases 2 and 3-1) evaluates the synthesis of information from conflicting arguments to reconstruct a coherent *Factual Basis*. **Statutory Symbolic Grounding** (Phases 3-2 and 3-3) tests the precision of legal provision citation and supports the mitigation of hallucination. **Logical & Numerical Inference** (Phases 4-1 and 4-3) captures System 2 reasoning in the exe-

cution of fixed algorithms and jurisdictional rules. **Constrained Legal Generation** (Phases 3-4 and 4-2) evaluates the generation of logically consistent judicial rationales and outcomes under strict procedural constraints.

3.3 Complexity Stratification and Evaluation Strategy

JurisBench evaluates legal reasoning under different levels of case complexity by using a three-tier stratification scheme: *Simple*, *Intermediate*, and *Complex*. Evaluation follows a simple-to-complex progression, in which models advance to higher-complexity subsets only after demonstrating adequate performance on simpler cases. This design reflects judicial practice in distinguishing difficult cases from complex cases and supports efficient identification of both baseline competence and the limits of expert-level reasoning.

The distinction among case complexity levels is grounded in established criteria for assessing adjudicative difficulty in judicial practice. These criteria encompass multiple factors that affect both factual complexity and the complexity of legal relations, including the difficulty of fact-finding and the structure of the legal relations involved. These factors are not arbitrarily defined; rather, they are abstracted from statutory provisions and judicial interpretations that explicitly recognize their effect on trial organization and adjudicative effort (Supreme People's Court of the PRC, 2022c; National People's Congress of the PRC, 2021).

On the basis of aggregated complexity scores derived from these normatively grounded indicators, the Jenks Natural Breaks Optimization method (Jenks, 1967) is used to partition cases into three difficulty levels with coherent legal characteristics and clear separation in adjudicative complexity. Detailed definitions of the indicators and scoring procedures are provided in Appendix E.

3.4 Evaluation Metrics

To assess model performance across the diverse cognitive demands of the judicial process, we categorize the evaluation metrics into four functional groups that cover all subtasks from initial filing to final judgment.

Classification and Exact Match. For the prediction of *Cause of Action* (including the representative disputes in Phase 1, as well as Phase 1-a and Phase 1-b), we use **Level-wise Top-3 Accuracy**. Under this metric, a prediction at a given hierarchical level is considered correct if the ground-truth label is included among the model's three highest-probability candidates. Given the hierarchical structure of the Chinese legal ontology, evaluation for a specific subcategory (level $l+1$) is conditioned on the ground truth of its parent category (level l). This design reflects the deductive reasoning process of judges and avoids penalizing valid high-level categorizations because of minor errors at the leaf level. For the analysis of *Appeal Jurisdiction* (Phase 4-3), we use **Exact Match** to ensure that the model correctly identifies the specific appellate court and its location.

Set-based Retrieval. For the prediction of *Legal Grounds*, which involves identifying both the relevant *Laws* (Phase 3-2) and the specific *Articles* (Phase 3-3), we use the **F1-score**. Because a single case typically depends on multiple statutory references, this metric balances the precision of cited provisions with the recall of all required legal grounds, thereby penalizing both the omission of critical laws and the hallucination of irrelevant ones.

Discretionary Numerical Tolerance. For the calculation of individual *Remedy* costs (Phase 4-1), we apply **$\pm$10% Tolerance Exact Match**. This metric reflects the legally sanctioned margin of judgment in Chinese civil law, under which compensation items such as "mental distress" or "reasonable expenses" are determined by locality-specific socioeconomic statistics and judicial discretion rather than by a unified national schedule (National People's Congress of the PRC, 2020; Supreme People's Court of the PRC, 2003). A prediction is considered correct if it falls within a 10% interval around the reference amount.

Semantic Generative Evaluation. For narrative and complex reasoning tasks, including the identification of *Focus of Disputes*, the extraction of *Factual Basis*, the generation of *Rationale of the Judgment*, and the determination of the final *Result of the Judgment* (Phases 2, 3-1, 3-4, and 4-2), surface-level n-gram metrics such as ROUGE (Lin, 2004) are insufficient. To address synonymy and logical sensitivity (e.g., "support" versus "dismiss"), we adopt a three-part evaluation strategy: **Embedding-based Similarity**, which calculates cosine similarity between vector representations; **LLM-as-a-Judge**, which uses a high-capability model to as-

Table 2: **Detailed task design and evaluation metrics for JurisBench.** We define subtasks across four judicial phases and specify their sequential input configurations, target outputs, and evaluation metrics.

Index	Task Description	Output	Judicial Capability	Metric
Phase 1: *Cause of Action* Prediction				
Input: Basic Case Info + Plaintiff's Argument				
1	MVTAL disputes, taxonomy-guided	*Cause of Action*	Semantic Taxonomy Alignment	Level-Acc.
1-a	Predict general 282 *Causes of Action*, taxonomy-guided	*Cause of Action*	Semantic Taxonomy Alignment	Level-Acc.
1-b	Predict general 282 *Causes of Action*, open-ended	*Cause of Action*	Semantic Taxonomy Alignment	Level-Acc.
Phase 2: *Focus of Disputes* Prediction				
Input: Basic Case Info + Plaintiff's Argument + Defendant's Argument + Cause of Action				
2	Evaluate each disputed issue	*Focus of Disputes*	Adversarial Factual Reconstruction	STS, EM
Phase 3: *Rationale of the Judgment* Prediction				
Input: Basic Case Info + Plaintiff's Argument + Defendant's Argument + Cause of Action + Facts Proved by the Court				
3-1	Extract useful *Factual Basis* for *Rationale of the Judgment*	*Factual Basis*	Adversarial Factual Reconstruction	STS
3-2	Retrieve case-relevant legal provisions	*Legal Grounds* (*Law*)	Statutory Symbolic Grounding	F1-score
3-3	Predict the exact legal provisions for *Rationale of the Judgment*	*Legal Grounds* (*Article*)	Statutory Symbolic Grounding	F1-score
3-4	Predict *Rationale of the Judgment* based on the litigation request	*Rationale of the Judgment*	Constrained Legal Generation	STS
Phase 4: *Result of the Judgment* Prediction				
Input: Basic Case Info + Plaintiff's Argument + Defendant's Argument + Cause of Action + Facts Proved by the Court + Rationale of the Judgment				
4-1	Calculate individual costs	*Remedy*	Logical & Numerical Inference	Tol-EM
4-2	Predict the individual *Result of the Judgment* based on the litigation request	*Result of the Judgment*	Constrained Legal Generation	STS
4-3	Analyze *Appeal Jurisdiction*	*Appeal Jurisdiction*	Logical & Numerical Inference	EM

Note: Level-Acc.: Level-wise Top-3 Accuracy; **STS**: Semantic Textual Similarity; **Tol-EM**: $\pm 10\%$ Tolerance Exact Match; **EM**: Exact Match; **F1**: F1-score (details are provided in Section 3.4).

sess factual consistency and logical coherence; and **Keyword-based Lexical Accuracy**, under which annotators with legal backgrounds identify a set of required professional keywords for each generative subtask (see Appendix G). An expert-led validation study confirmed that these automated proxies maintain high correlation with manual ratings provided by professional judges (see Appendix G). For these phases, the primary results in the main text are computed by using the first two metrics, while a complementary evaluation based on professional keyword recovery is provided in Appendix F.4.

4 Experiment

4.1 Experimental Setup

We evaluate 13 representative LLMs, encompassing state-of-the-art Multilingual models (Claude Sonnet 4.5 (Anthropic, 2025), GPT-4o Mini (OpenAI, 2024), Gemini-3 Pro Preview (Google DeepMind, 2025), Llama 3.3 70B Instruct (Meta AI, 2024), and Grok 4 (xAI, 2025)) and Chinese-Oriented models (DeepSeek-R1 (DeepSeek-AI), Doubao-1.5-Pro (ByteDance, 2025), Kimi-K2 (Kimi Team, 2025), Qwen3-Instruct (Qwen Team, 2025), and GLM-4.5 (GLM-4.5 Team, 2025)). Additionally, 3 Legal-Specific models are included to assess the impact of domain-specific fine-tuning (see Appendix F.3).

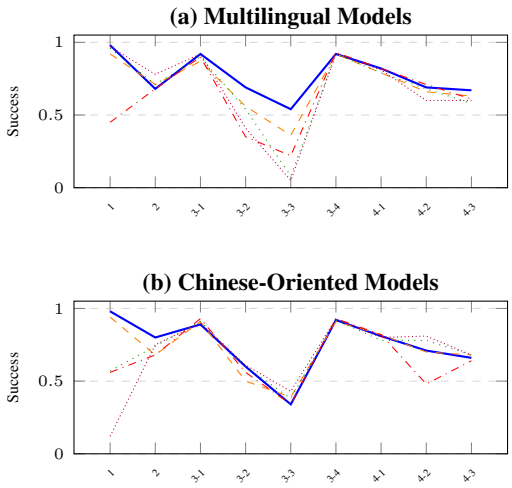


Figure 3: Parallel coordinates plot visualizing performance decay (oracle-assisted evaluations). The longitudinal axis represents accuracy across sequential pipeline phases. Each line represents a LLM under test; the blue lines represent Gemini-3 Pro and Qwen3-Instruct respectively.

4.2 Methodology & Main Results

Workflow-Oriented Modular Evaluation To precisely measure model capabilities at each judicial stage, we first adopt a *modular evaluation* strategy. For each pipeline phase, the model is provided with ground-truth (*gold*) inputs from all

Table 3: Experimental results (zero-shot) of 10 LLMs (*Simple* difficulty subset). **Bold** and underline indicate the best and second-best performance in each column, respectively. Cell background colors represent a performance heatmap, transitioning from red (lower performance) to green (higher performance). The "Product" column represents per-mille (‰) success rate with one decimal place.

		Phase 1	Phase 2	Phase 3				Phase 4			
Group	**Model**	1	2	3-1	3-2	3-3	3-4	4-1	4-2	4-3	**Product**
Multilingual	Gemini-3 Pro	**0.982**	0.684	0.917	**0.689**	**0.535**	0.919	**0.820**	0.693	0.669	**79.3‰**
	Claude 4.5	0.923	0.711	0.868	0.561	0.355	0.921	0.790	0.661	0.626	34.2‰
	GPT-4o-mini	0.957	0.709	0.901	0.542	0.083	0.919	0.794	0.681	0.576	7.9‰
	Grok 4	0.447	0.681	0.901	0.347	0.224	0.926	0.817	0.711	0.606	6.9‰
	Llama 3.3 70B	0.970	0.782	0.917	0.413	0.051	0.923	0.808	0.604	0.598	3.9‰
Chinese-Oriented	Qwen3-Instruct	**0.982**	**0.798**	0.892	0.602	0.336	0.924	0.810	0.710	0.660	**49.5‰**
	Doubao-1.5-Pro	0.937	0.679	0.909	0.502	0.387	0.909	0.818	0.700	0.679	39.7‰
	GLM-4.5	0.571	0.738	0.912	0.604	0.383	0.922	0.776	0.776	0.675	33.3‰
	DeepSeek-R1	0.556	0.683	**0.927**	0.564	0.344	0.927	0.815	0.480	0.640	15.9‰
	Kimi-K2	0.117	0.750	0.904	0.607	0.427	**0.930**	0.800	**0.807**	**0.684**	8.4‰

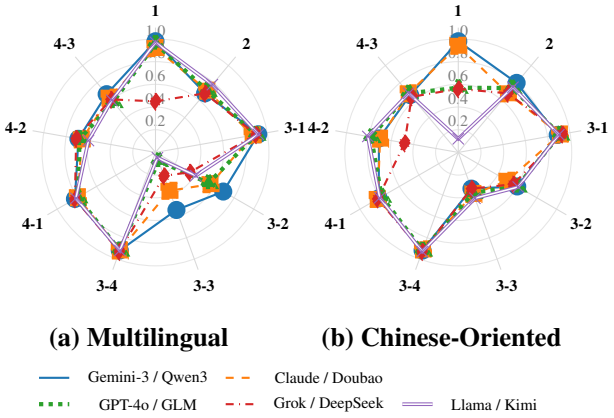
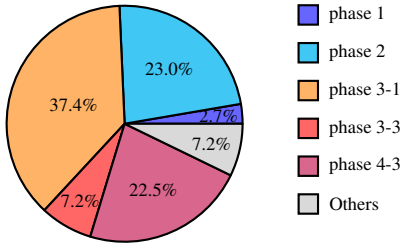
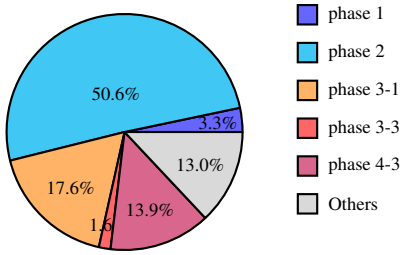


Figure 4: Capability radar charts across 9 sub-tasks of the 4 phases (oracle-assisted evaluations). Concentric rings indicate accuracy levels (0.2–1.0).



(a) Qwen3-Instruct (First Failure Points)



(b) Gemini-3 Pro (First Failure Points)

Figure 5: Distribution of the first point of failure in the unassisted workflow.

preceding stages (e.g., the true *Cause of Action* is provided as a premise for *Focus of Disputes* identification). Detailed results are shown in Table 3. This "Oracle-assisted" approach ensures that performance failures at a later stage are not confounded by reasoning errors inherited from earlier phases, allowing for a granular diagnosis of specific capability bottlenecks (visualized in Figure 3 and 4).

Unassisted Simulation and Systemic Reliability To investigate "true" end-to-end reliability—where errors are unassisted and cumulative—we transition from oracle assistance to a sequential, *unassisted simulation* on top-performing models (Gemini-3 Pro and Qwen3-Instruct). We enforce a *strict dependency chain*: for each individual case, if a model yields a score of 0 at any stage, it triggers a **Terminal Pipeline Failure**, rendering all downstream reasoning moot (see Figure 6).

To reflect the overall coherence, we define the **Product** metric (‰) as the joint success rate. While the modular Product (Table 3) represents an *idealized upper bound*, the unassisted simulation tracks the *Effective Systemic Reliability* and the distribution of **First Failure Points** (Figure 5) to identify where the reasoning chain initially fractures.

Prompting Strategy and Scope All primary results reported in the main text are obtained under a **zero-shot** setting to assess the models' intrinsic legal reasoning capabilities. The impact of in-context learning via **one-shot** demonstrations, as well as the evaluation of Phase 1-a and 1-b, are detailed in Appendix F.

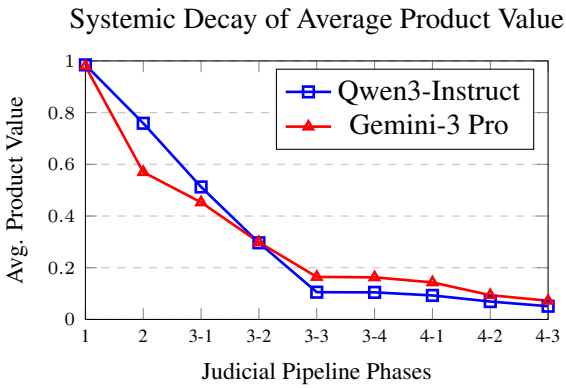


Figure 6: Systemic decay of the **Average Product** across the unassisted workflow. Plotted values represent the mean joint success rate calculated *only* for the subset of cases that have successfully traversed all preceding phases without a terminal failure. This visualization highlights the catastrophic impact of cascading errors on overall systemic reliability.

4.3 Analysis

Our analysis of the experimental results yields the following core findings:

The Bottleneck of Statutory Symbolic Grounding While most models exhibit robust performance in *Semantic Alignment* (Phase 1, avg. >0.9) and *Factual Reconstruction* (Phase 3-1, avg. >0.9), they struggle significantly with *Symbolic Grounding*. The capability footprints in Figure 4 reveal a universal "indentation" at Phase 3-3 (*Legal Grounds-Article Prediction*), where even Gemini-3 Pro achieves only 0.535. The parallel coordinates (Figure 3) present a stark "competence gap": LLMs excel at summarizing narratives but lack the "legal fidelity" required to anchor facts to specific statutory provisions. This acts as a **hard filter**: once a model fails this precise grounding, the subsequent accuracy in the pipeline decays drastically.

Modular Success vs. Systemic Fragility A striking discrepancy exists between modular scores and unassisted reliability. As shown in the simulation (Figure 6), systemic performance exhibits a catastrophic cascade effect. By the final phase (4-3), even the best-performing models achieve a **conditional pipeline success** of merely **5.12%** (Qwen3) and **7.22%** (Gemini-3 Pro).

Critically, these values represent a *conditional average*—calculated only from the subset of cases that have not yet encountered a terminal failure in preceding stages. As illustrated by the "reasoning fractures" in Figure 5, the vast majority of cases are filtered out much earlier; for instance, 50.6% of Gemini-3 Pro cases fracture at Phase 2. This implies that the *Effective Systemic Reliability*, factoring in the cumulative **Survival Rate**, is significantly lower than even these marginal percentages suggest. This evidence proves that judicial reasoning behaves as a "weakest-link" system: a single failure in early-stage information synthesis (as seen in Qwen3's 37.4% fracture at Phase 3-1) effectively nullifies any linguistic fluency achieved in later generative stages.

Thinking-Heavy Models vs. General-Purpose Stability A nuanced trade-off emerges between specialized reasoning and end-to-end stability. As shown in Table 3, while DeepSeek-R1 exhibits the highest stability in the initial extraction of the *Factual Basis* (Phase 3-1, 0.927), its final **Product** metric remains relatively low (15.9‰), suggesting that general-purpose reinforcement learning for reasoning ("thinking" traces) does not automatically translate into the procedural rigor required for a multi-stage judicial workflow.

Furthermore, we observe a "specialization split" among Chinese-oriented models: for instance, Kimi-K2 shows the weakest performance in initial *Cause of Action* alignment (Phase 1, 0.117) but demonstrates a superior ability in final *Result of the Judgment* determination (Phase 4-2, 0.807). Despite these localized strengths, Gemini-3 Pro maintains a dominant lead in the final **Product** success rate (79.3‰). This highlights a critical need for future legal AI to bridge the gap between "thinking" prowess and "procedural" consistency.

5 Conclusion

JurisBench introduces a workflow-aligned benchmark for evaluating LLMs in professional legal case-processing scenarios, emphasizing domain-specific depth rather than broad but shallow task coverage. By evaluating model performance across interdependent judicial stages, the benchmark shows that strong results on isolated legal subtasks do not necessarily translate into coherent end-to-end case handling: early-stage errors frequently propagate, and precise statutory grounding remains a persistent bottleneck. Although current models perform unsatisfactorily under this evaluation, these results highlight the diagnostic value of JurisBench in revealing failure modes that are often obscured by exam-style or surface-level benchmarks.

Beyond the legal domain, JurisBench also illustrates a broader benchmark design perspective for high-stakes professional settings. Its workflow-centric formulation may inform benchmark construction in other specialized domains, such as healthcare, financial regulation, and engineering decision-making, where reliability across the full decision process is more important than performance on isolated subtasks.

Limitations

JurisBench has several limitations. First, **domain coverage** is restricted to a single dispute type, motor vehicle traffic accident liability. Although the adjudicative workflow from *Cause of Action* to *Focus of Disputes*, *Rationale*, and *Result* is broadly representative of civil adjudication, generalization to other civil domains is not empirically validated. Second, **metric limitations** remain. While our embedding-based and LLM-assisted metrics correlate well with expert judgments, automated evaluation cannot fully capture the full nuance and rigor of judicial reasoning. Third, **benchmark staleness and contamination** are potential concerns. As a fixed benchmark, JurisBench may be exposed to data contamination from future model training, and evolving judicial interpretations may require periodic updates to preserve legal validity. Finally, although we study both modular and unassisted pipeline settings, we do not evaluate fully autonomous agentic workflows or richer interactive reasoning architectures, which remain a promising direction for future research.

Ethics Statement & Potential Risks

Preventing Misinterpretation. We explicitly state that JurisBench is a *diagnostic tool* for identifying reasoning fractures in LLMs, not a certification for autonomous adjudication. Our findings, particularly the low "Product" success rates, underscore that current LLMs are not ready for high-stakes judicial decision-making. High benchmark scores should not be equated with the professional and moral judgment of a human judge.

Bias and Privacy. As JurisBench is derived from authentic judicial documents, it may reflect historical systemic biases. Models evaluated on this data might inadvertently reinforce these biases; thus, results should not be viewed as absolute legal "correctness." Regarding privacy, although we implemented rigorous anonymization for names and identifiers, the unique combination of case facts in open datasets poses a theoretical risk of re-identification. We urge the community to use this data strictly for research purposes and in compliance with global data protection standards (e.g., GDPR or local equivalents).

Jurisdictional Scope. JurisBench is tailored to the Chinese civil law system. Users should exercise extreme caution when applying its evaluation logic to common law jurisdictions or other legal frameworks, as reasoning patterns and procedural requirements differ significantly.

Acknowledgments

We would like to express our sincere gratitude to the courts in Beijing and Shandong for their invaluable support and assistance throughout this research. We are especially thankful to the judges who provided insightful and practice-oriented suggestions during the design of the benchmark, and who also contributed directly to the evaluation process. Their professional expertise greatly enhanced the realism and rigor of the benchmark.

We are also deeply grateful to the faculty members who offered constructive feedback and guidance on the benchmark design at various stages of this work. Their suggestions were instrumental in shaping the overall framework and evaluation methodology.

Finally, we thank the students who devoted substantial effort to data construction, annotation, and verification. Their careful and collaborative work was essential to ensuring the quality and reliability of the dataset and the benchmark as a whole.

References

Anthropic. 2025. Claude sonnet 4.5 announcement. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2026-01-05.

Kevin D. Ashley. 2017. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, Cambridge.

Beijing Chaoyang District People's Court. 2025. Deepening the reform of separating complex and simple civil litigation procedures (reform case no. 180). Supreme People's Court Website, Accessed: 2025-11-19.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and

1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

ByteDance. 2025. Doubao-1.5-pro model documentation. https://seed.bytedance.com/zh/special/doubao_1_5_pro. Accessed: 2026-01-05.

Ilias Chalkidis, Michalis Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Ilias Chalkidis, Nicolas Garneau, Cătălina Goanță, Daniel Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535.

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.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models to domains via reading comprehension. *arXiv preprint arXiv:2309.09530*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and 1 others. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

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

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

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2023. Laiw: A chinese legal large language models benchmark. *arXiv preprint arXiv:2310.05620*.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning.

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.

Colin Doyle and Aaron D Tucker. 2025. If you give an llm a legal practice guide. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 194–205.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, and 1 others. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864*.

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

GLM-4.5 Team. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Google DeepMind. 2025. Gemini 3 pro. https://deepmind.google/models/gemini/pro/. Accessed: 2026-01-05.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Sazzad Hossain, Touhidul Alam Seyam, Avijit Chowdhury, Munis Xamidov, Rajib Ghose, and Abhijit Pathak. 2025. Fine-tuning llama 2 interference: a comparative study of language implementations for optimal efficiency. *arXiv preprint arXiv:2502.01651*.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

George F. Jenks. 1967. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7:186–190.

Paul Kassianik, Baturay Saglam, Alexander Chen, Blaine Nelson, Anu Vellore, Massimo Aufiero, Fraser Burch, Dhruv Kedia, Avi Zohary, Sajana Weerawardhena, and 1 others. 2025. Llama-3.1-foundationai-securityllm-base-8b technical report. *arXiv preprint arXiv:2504.21039*.

Kimi Team. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Katherine Lee, A Feder Cooper, James Grimmelmann, and Daphne Ippolito. 2023. Ai and law: The next generation. *Available at SSRN 4580739*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, and 1 others. 2024a. Legalagentbench: Evaluating llm agents in legal domain. *arXiv preprint arXiv:2412.17259*.

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

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

Dong Liu and Yuan Gao. 2015. Shanghai courts pioneer case weight coefficient assessment. *Wen Hui Bao*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Meta AI. 2024. Introducing meta llama 3. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2026-01-05.

National People's Congress of the PRC. 2020. Civil code of the people's republic of china, article 1179.

National People's Congress of the PRC. 2021. Civil procedure law of the people's republic of china (article 160). Legislative Law.

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

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. Multilegalpile: A 689gb multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094.

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

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2026-01-05.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Eric A Posner and Shivam Saran. 2025. Judge ai: Assessing large language models in judicial decision-making. *University of Chicago Coase-Sandor Institute for Law & Economics Research Paper*, (2503).

Philippe Prince Tritto and Hiram Ponce. 2025. Assessing ai-generated legal reasoning: A benchmark for legal text quality from literature review. In *Mexican Congress on Artificial Intelligence*, pages 54–68. Springer.

Haosheng Qian, Yixing Fan, Jiafeng Guo, Ruqing Zhang, Qi Chen, Dawei Yin, and Xueqi Cheng. 2025. Vericite: Towards reliable citations in retrieval-augmented generation via rigorous verification. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 47–54.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E Ho, and Joel Niklaus. 2024. One law, many languages: Benchmarking multilingual legal reasoning for judicial support. In *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR): Harnessing Momentum for Science. https://openreview.net/forum*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Shandong High People's Court. 2025. Characteristics of road traffic disputes and governance suggestions.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*.

Standing Committee of the National People's Congress of the PRC. 2023. Civil procedure law of the people's republic of china (2023 amendment). Adopted by the Standing Committee of the National People's Congress; effective January 1, 2024.

State Council of the People's Republic of China. 2007. Measures for the payment of litigation costs. Administrative regulations governing litigation fees; as amended.

Supreme People's Court of the PRC. China judgments online. https://wenshu.court.gov.cn/.

Supreme People's Court of the PRC. 2003. Interpretation on several issues concerning the application of law in the trial of personal injury compensation cases.

Supreme People's Court of the PRC. 2016a. 2015 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2016b. Provisions of the supreme people's court on the drafting of judicial documents. Judicial provisions regulating the structure and content of court judgments.

Supreme People's Court of the PRC. 2016c. Work report of the supreme people's court – delivered at the fourth session of the 12th national people's congress on march 13, 2016. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2017a. 2016 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2017b. Work report of the supreme people's court – delivered at the fifth session of the 12th national people's congress on march 12, 2017. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2018a. 2017 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2018b. Work report of the supreme people's court – delivered at the first session of the 13th national people's congress on march 9, 2018. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2019a. 2018 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2019b. Provisions of the supreme people's court on evidence in civil litigation. Judicial provisions governing evidence in civil litigation.

Supreme People's Court of the PRC. 2019c. Work report of the supreme people's court – delivered at the second session of the 13th national people's congress on march 12, 2019. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2020a. 2019 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2020b. Provisions on the causes of action of civil cases. Judicial provisions establishing the classification system of civil causes of action.

Supreme People's Court of the PRC. 2020c. Work report of the supreme people's court – delivered at the third session of the 13th national people's congress on may 25, 2020. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2021a. 2020 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2021b. Work report of the supreme people's court – delivered at the fourth session of the 13th national people's congress on march 8, 2021. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2022a. 2021 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2022b. Interpretation of the supreme people's court on the application of the civil procedure law of the people's republic of china. Judicial interpretation issued by the Supreme People's Court.

Supreme People's Court of the PRC. 2022c. Interpretation of the supreme people's court on the application of the civil procedure law of the people's republic of china (article 257). Judicial Interpretation.

Supreme People's Court of the PRC. 2022d. Work report of the supreme people's court – delivered at the fifth session of the 13th national people's congress on march 8, 2022. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2023a. 2022 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2023b. Work report of the supreme people's court – delivered at the first session of the 14th national people's congress on march 7, 2023. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2024a. 2023 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2024b. Work report of the supreme people's court – delivered at the second session of the 14th national people's congress on march 8, 2024. Accessed: 2025-11-01.

Supreme People's Court of the PRC. 2025. 2024 national judicial statistical bulletin of chinese courts. Accessed: 2025-11-01.

Harry Surden. 2018. Artificial intelligence and law: An overview. *Ga. St. UL Rev.*, 35:1305.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Robert Walters and Marko Novak. 2021. Artificial intelligence and law. In *Cyber security, artificial intelligence, data protection & the law*, pages 39–69. Springer.

Lan Wang and Sufang Qiu. 2019. Measurement of judges' workload: Econometric models and sichuan experience. *Journal of Shanghai Jiao Tong University (Philosophy and Social Sciences)*, 27(6):61–73.

xAI. 2025. Grok 4. https://x.ai/news/grok-4. Accessed: 2026-01-05.

Shunyu Yao, Qingqing Ke, Qiwei Wang, Kangtong Li, and Jie Hu. 2024. Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, pages 108–112.

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

Kepu Zhang, Weijie Yu, Sunhao Dai, and Jun Xu. 2024. Citalaw: Enhancing llm with citations in legal domain. *arXiv preprint arXiv:2412.14556*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D Manning, Peter Henderson, and Daniel E Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 169–193.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

A Increasing Caseload in China

Data extracted from the annual *Work Report of the Supreme People's Court* and the *National Judicial Statistical Bulletin of Chinese Courts* reveals a consistent and significant upward trend in the number of cases handled by Chinese courts over the past decade. This growth has been accompanied by severe pressure arising from the "judge–case conflict," namely the structural imbalance between the number of judges and the rapidly increasing caseload.

From 2015 to 2019, the number of first-instance civil cases rose steadily from 6.228 million to 9.393 million, reflecting the expansion of disputes in civil domains such as family, labor, and consumer protection (Supreme People's Court of the PRC, 2016c, 2017b, 2018b, 2019c, 2020c).

Table 4: First-instance Civil Cases in China (2015–2019).

Year	Number of Cases (10,000 cases)
2015	62.28
2016	67.38
2017	82.95
2018	90.17
2019	93.93

Note: Data for years after 2019 is not available in the Work Report of the Supreme People's Court. Data labeled as "first-instance civil cases" in the National Judicial Statistical Bulletin of Chinese Courts actually refers to civil and commercial cases and therefore cannot be used as a direct supplement.

For first-instance civil and commercial cases, which cover a broader range of business-related disputes, the increase was even more pronounced. The caseload grew from 9.575 million in 2015 to 18.237 million in 2024, nearly doubling over a ten-year period (Supreme People's Court of the PRC, 2016a, 2017a, 2018a, 2019a, 2020a, 2021a, 2022a, 2023a, 2024a, 2025).

The rapid expansion of case volume has directly translated into a heavier workload for judges. The annual average number of cases handled per judge increased from 187 in 2017 to 357 in 2023. Intermediate statistics indicate a persistent upward trajectory, with averages of 225 cases in 2020, 238 in 2021, and 242 in 2022 (Supreme People's Court of the PRC, 2021b, 2022d, 2023b, 2024b).

The Supreme People's Court has explicitly acknowledged that "the judge–case" conflict has become increasingly prominent" underscoring the mounting pressure faced by the judicial workforce (Supreme People's Court of the PRC, 2024b).

Table 5: First-instance Civil and Commercial Cases in China (2015–2024)

Year	Number of Cases (10,000 cases)
2015	95.75
2016	107.64
2017	116.51
2018	124.35
2019	139.30
2020	133.06
2021	157.46
2022	161.14
2023	174.77
2024	182.37

Table 6: Annual Average Cases Handled per Judge in China (2017, 2020–2023).

Year	Average Cases per Judge
2017	187
2020	225
2021	238
2022	242
2023	357

B Data Construction and Filtering

Table 7 presents detailed statistics of field-level completeness after structured information extraction from the raw legal documents. Each row corresponds to a specific field required for constructing or characterizing cases in JurisBench, and the missing rate reflects the proportion of documents in which the corresponding field cannot be reliably extracted. All raw legal documents used in this study are sourced from judicial decisions published in 2024 or later.

The results show that the overall completeness of extracted legal documents is limited. While several meta-level attributes, such as case name and *Cause of Action*, exhibit negligible missing rates, many core litigation fields suffer from substantial absence. In particular, high-level *Rationale of the Judgment* components, including *Focus of Disputes* and *Facts Proved by the Court*, are missing in a large portion of documents, reflecting both the heterogeneity of judicial writing styles and the inherent difficulty of automatically identifying

fine-grained legal reasoning elements. As a consequence, only a relatively small subset of documents contains all required fields and can be considered fully usable for benchmark evaluation.

It is important to note that several fields listed under *Basic Case Info*, such as *Trial Procedure*, *Case Number*, and *Closing Date*, are not directly used as inputs or targets in JurisBench tasks. Nevertheless, we treat these attributes as mandatory metadata and exclude documents in which they are missing. This design choice is motivated by long-term benchmark reliability: missing basic case information may introduce inconsistencies in dataset statistics, hinder reproducibility, and complicate future extensions involving temporal analysis, procedural stratification, or cross-case aggregation. By enforcing completeness even for non-task fields, we ensure that all retained cases are well-formed, uniquely identifiable, and suitable for systematic statistical analysis.

Overall, this filtering process results in $N_{usable} = 3{,}204$ fully structured cases, yielding a ratio of approximately 19.13 incomplete documents for every usable one. We selected the 1,000 cases via uniform random sampling from the full set ($N_{usable} = 3{,}204$) because of the high cost of expert verification required to establish a rigorous "Gold Standard" for pipeline evaluation. The complexity stratification (Simple/Intermediate/Complex) was performed exclusively on the 1,000 sampled cases based on their verified ground truth. Although this substantially reduces the dataset size, it provides a strong guarantee on data integrity and supports reliable, interpretable, and extensible evaluation of large language models in the legal domain.

C Normative Foundations of the JurisBench Pipeline

The JurisBench pipeline is normatively grounded in the procedural framework of Chinese civil litigation. All directed arrows in the benchmark correspond to legally mandated or institutionally established dependencies in judicial practice. Rather than representing transitions between isolated case elements, these arrows denote ordered transitions between sub-benchmarks, each of which operates on a structured and composite input set. We distinguish between **inter-benchmark (macro) transitions**, which connect adjacent sub-benchmarks in the pipeline, and **intra-benchmark (micro) transitions**, which decompose reasoning steps within a single sub-benchmark.

C.1 Inter-Benchmark (Macro) Transitions

Phase 1 (*Cause of Action*) → Phase 2 (*Focus of Disputes*). The transition from Phase 1 to Phase 2 reflects the procedural dependency between legal characterization and dispute identification in Chinese civil litigation. Phase 1 determines the *Cause of Action* based on a composite input consisting of *Basic Case Info* and *Plaintiff's Argument*, corresponding to the statutory case-filing stage. Under the Civil Procedure Law of the People's Republic of China, a civil action must be initiated through a written complaint specifying claims, facts, and reasons, which constitute the legally required basis for determining the legal nature of the dispute at docketing (Standing Committee of the National People's Congress of the PRC, 2023). This determination governs subsequent trial organization and adjudication, including judicial division assignment.

Phase 2 operates on an expanded input set that includes *Basic Case Info*, *Plaintiff's Argument*, *Defendant's Argument*, and the *Cause of Action* output from Phase 1, and evaluates the identification of *Focus of Disputes* under adversarial conditions. *Judicial Interpretations* authorize courts to summarize dispute foci during pretrial proceedings and require adjudication to proceed around such foci (Supreme People's Court of the PRC, 2022b, 2019b). Since dispute identification must be framed within the legally established nature of the case, the arrow from Phase 1 to Phase 2 denotes the normative requirement that *Focus of Disputes* identification be conditioned on a prior determination of *Cause of Action*, rather than treated as an unconstrained issue-extraction task.

Phase 2 (*Focus of Disputes*) → Phase 3 (*Rationale of the Judgment*). The transition from Phase 2 to Phase 3 corresponds to the statutory ordering between dispute identification and *Rationale of the Judgment*. Phase 2 produces *Focus of Disputes* based on a composite input integrating *Basic Case Info*, *Plaintiff's Argument*, *Defendant's Argument*, and *Cause of Action*, reflecting the adversarial identification of contested issues. Civil adjudication is required to proceed around disputed facts and issues, and judgments must explicitly address these disputes through reasoned analysis (Standing Committee of the National People's Congress of

Table 7: **Data Filtering Ratio Statistics.** This table reports field-level missing statistics during the structured extraction of raw legal documents, illustrating substantial variation in completeness across different types of information. After enforcing strict completeness requirements, only 3,204 cases contain all required fields and are retained as usable data.

Category	Field Name	Missing	Total	Missing Rate (%)
Meta Information	*Meta Information*	0	64,504	0.00
	Party	11,736	64,504	18.19
Core Litigation	*Plaintiff's Argument*	812	64,504	1.26
	Defendant's Argument	8,573	64,504	13.29
	Focus of Disputes	55,480	64,504	86.01
	Facts Proved by the Court	18,474	64,504	28.64
	Rationale of the Judgement	3,894	64,504	6.04
	Result of the Judgement	1,068	64,504	1.66
Basic Case Info	*Court of Acceptance*	505	64,504	0.78
	Trial Procedure	107	64,504	0.17
	Case Name	0	64,504	0.00
	Case Number	1,253	64,504	1.94
	Cause of Action	0	64,504	0.00
	Closing Date	845	64,504	1.31

Note: Ratio of incomplete to usable documents ($N_{usable} = 3,204$) is 19.13:1.

the PRC, 2023).

Phase 3 evaluates *Rationale of the Judgment* under a further expanded information set consisting of *Basic Case Info*, *Plaintiff's Argument*, *Defendant's Argument*, *Cause of Action*, and *Facts Proved by the Court*. Although *Focus of Disputes* is not treated as an explicit input variable, it is institutionally embedded in the task design as a binding contextual constraint. The arrow from Phase 2 to Phase 3 therefore represents the legal requirement that *Rationale of the Judgment* respond to previously identified dispute foci rather than being generated independently of the adversarial issue structure.

Phase 3 (*Rationale of the Judgment*) → Phase 4 (*Result of the Judgment*). The transition from Phase 3 to Phase 4 reflects the formal structure of civil judgments. Phase 3 assesses the model's ability to generate *Rationale of the Judgment* grounded in established facts and applicable law, based on a composite input consisting of *Basic Case Info*, *Plaintiff's Argument*, *Defendant's Argument*, *Cause of Action*, and *Facts Proved by the Court*.

Phase 4 operates on an augmented input set that incorporates the *Rationale of the Judgment* output of Phase 3 in addition to *Basic Case Info*, *Plaintiff's Argument*, *Defendant's Argument*, *Cause of Action*, and *Facts Proved by the Court*, and evaluates whether the model can derive *Result of the Judgment* that is procedurally and substantively consistent with the preceding reasoning. Statutory provisions require civil judgments to include both the reasoning and the adjudicative outcome, rendering *Result of the Judgment* the logical culmination of *Rationale of the Judgment* (Standing Committee of the National People's Congress of the PRC, 2023). Accordingly, the arrow from Phase 3 to Phase 4 denotes the legally mandated dependency between *Rationale of the Judgment* and *Result of the Judgment*.

C.2 Intra-Benchmark (Micro) Transitions

Internal Structure of Phase 1 (*Cause of Action*). Within Phase 1, the hierarchical progression across levels of *Cause of Action* reflects the multi-level classification system established by the Provisions on the *Causes of Action* of Civil Cases (Supreme People's Court of the PRC, 2020b). This internal structure supports standardized case filing and adjudication by progressively refining the legal characterization of disputes, while remaining anchored in the same composite input of *Basic Case Info* and *Plaintiff's Argument*.

Internal Structure of Phase 2 (*Focus of Disputes*). Within Phase 2, *Focus of Disputes* identification presupposes the joint consideration of *Plaintiff's Argument* and *Defendant's Argument* as part of its composite input. This reflects the adversarial structure of civil litigation, under which courts must hear arguments from both parties before determin-

ing contested issues (Standing Committee of the National People's Congress of the PRC, 2023). The internal reasoning of Phase 2 therefore evaluates the synthesis and reconciliation of competing factual claims rather than unilateral issue extraction.

Internal Structure of Phase 3 (*Rationale of the Judgment*). Phase 3 decomposes *Rationale of the Judgment* into multiple normative steps, including grounding in *Facts Proved by the Court*, application of applicable law, citation of specific legal Articles, and synthesis of the written rationale. This internal decomposition is normatively supported by statutory principles requiring adjudication to be based on established facts and governed by law, as well as judicial drafting rules mandating explicit articulation of facts, *Legal Grounds*, and reasoning with accurate statutory citations (Standing Committee of the National People's Congress of the PRC, 2023; Supreme People's Court of the PRC, 2016b).

Internal Structure of Phase 4 (*Result of the Judgment*). Within Phase 4, the determination of litigation costs precedes the assessment of appeal jurisdiction, reflecting auxiliary procedural regulations governing litigation fees and appellate rights (State Council of the People's Republic of China, 2007; Standing Committee of the National People's Congress of the PRC, 2023). These internal steps jointly constitute the adjudicative outcome and are integrated into the finalized *Result of the Judgment* in accordance with statutory rules on the composition and effectiveness of civil judgments (Standing Committee of the National People's Congress of the PRC, 2023).

D Semantic Diversity and Dataset Distribution

To evaluate the representational capacity and semantic diversity of the **JurisBench** dataset, we performed a high-dimensional feature analysis followed by manifold learning-based visualization. Specifically, we first extracted representative textual features from the judicial documents using a *Term Frequency-Inverse Document Frequency* (TF-IDF) vectorization scheme (Salton and Buckley, 1988). We incorporated both unigrams and bigrams to capture fine-grained legal terminology and structural markers. Given the high dimensionality of the resulting feature space, we applied Principal Component Analysis (PCA) to reduce the latent space to 50 dimensions, followed by t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) for non-linear dimensionality reduction into a two-dimensional visual manifold.

The semantic distribution across the three difficulty tiers—Simple ($n = 394$), Intermediate ($n = 394$), and Complex ($n = 212$)—is illustrated in Figure 7. The visualization reveals an extensive and relatively uniform distribution across the latent space, avoiding narrow or isolated clusters. This suggests that **JurisBench** covers a broad and diverse semantic landscape within the legal domain, rather than being restricted to a few repetitive scenarios.

Quantitatively, our diversity analysis yields an average pairwise distance of 1.3603 ($\sigma = 0.0631$), which is remarkably close to the theoretical maximum distance of 1.414 in a normalized TF-IDF vector space. This high degree of sparsity and semantic spread indicates that the dataset possesses low redundancy and high task variance. Notably, the intermingling of simple, intermediate, and complex cases in the t-SNE plot demonstrates that case complexity in our benchmark is driven by procedural depth and reasoning logic rather than mere lexical variations or superficial keyword distributions, thereby validating the structural integrity of our difficulty stratification.

E Complexity Indicators and Stratification Procedure

The complexity stratification of JurisBench test cases is grounded in both statutory standards and empirical studies on case difficulty in Chinese judicial practice. Each case is quantitatively evaluated along five core dimensions: *Parties* Involved, *Claims*, *Focus of Disputes*, *Rationale of the Judgment*, and *Result of the Judgment*.

In the *Parties* Involved dimension, multi-party participation is treated as a primary driver of complexity, as it often entails intertwined legal relationships and challenges in liability allocation (Supreme People's Court of the PRC, 2022c; Shandong High People's Court, 2025; Beijing Chaoyang District People's Court, 2025). In the *Claims* and *Focus of Disputes* dimensions, compound claims and intensive disputes—such as disagreements over appraisal conclusions or jurisdictional objections—are indicative of increased trial difficulty. This aligns with procedural standards that restrict simplified procedures to cases with clear rights and limited disputes (National People's Congress of the

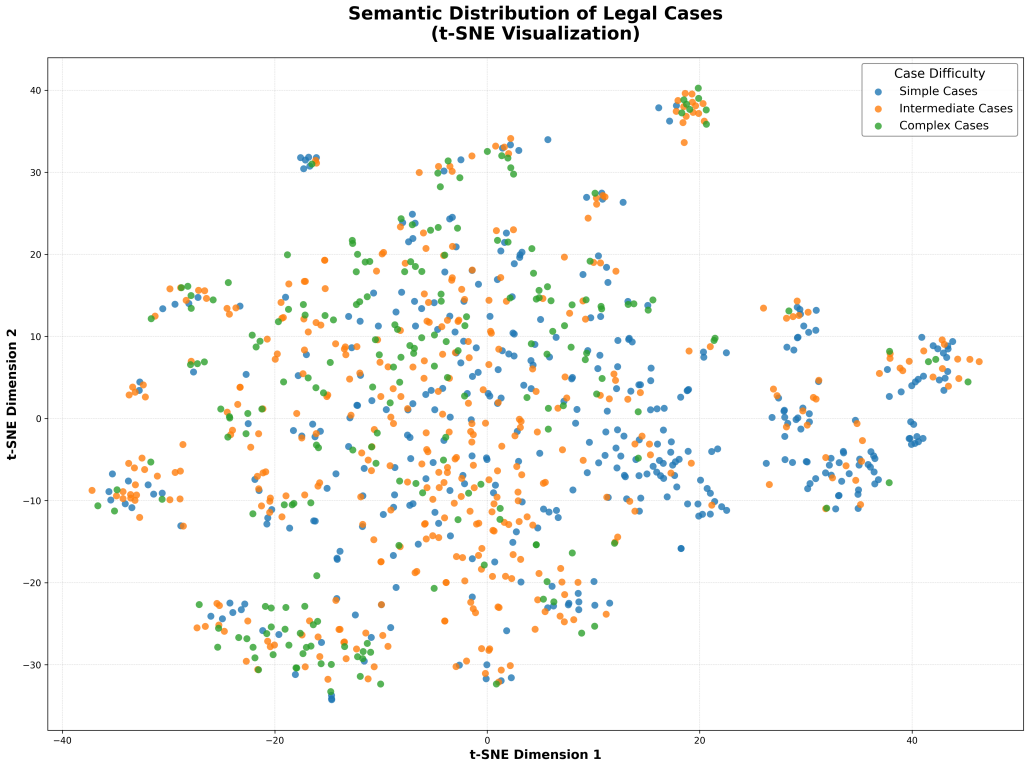


Figure 7: Semantic distribution of legal cases in JurisBench. The t-SNE visualization shows the high-dimensional embedding space of cases across three difficulty levels. The broad spread and high average pairwise distance (1.3603) reflect the semantic diversity and low redundancy of the dataset.

PRC, 2021).

In the *Rationale of the Judgment* and *Result of the Judgment* dimensions, the length and structural depth of judgment documents are used as proxy variables for judicial effort and reasoning complexity. Prior empirical studies and case-weight reform practices have shown that more extensive reasoning is strongly correlated with increased fact-finding difficulty and legal analysis depth (Liu and Gao, 2015; Wang and Qiu, 2019).

After scoring cases across all dimensions, the Jenks Natural Breaks Optimization method (Jenks, 1967) is applied to identify two optimal breakpoints in the complexity score distribution. This method minimizes within-group variance while maximizing between-group variance, enabling an objective partition of cases into *Simple*, *Intermediate*, and *Complex* subsets with coherent legal characteristics and clearly differentiated difficulty levels.

F Additional Experiments

F.1 General Legal Taxonomy Alignment (Phase 1-a & 1-b)

To verify the generalizability of LLMs across the full spectrum of Chinese civil law, we evaluated 10 representative models on a broad dataset comprising 2,820 cases across 282 *Causes of Action*. This experiment investigates the models' performance when provided with **hierarchical taxonomy guidance** (Phase 1-a) versus their purely **parameterized recall** in a open-ended setting (Phase 1-b).

Table 8: Results of General Legal Taxonomy Alignment on 282 *Causes of Action* (2,820 cases). Values in **bold** and underline indicate the best and second-best performance per column.

Model	Phase 1-a ↑	Phase 1-b ↑	Δ (Drop %)
Multilingual Models			
Gemini-3 Pro	**0.7128**	**0.3430**	-51.88%
Claude 4.5	0.6071	0.0699	-88.49%
GPT-4o-mini	0.5032	0.0011	-99.78%
Grok 4	0.6114	0.0096	-98.43%
Llama 3.3 70B	0.6233	0.0024	-99.62%
Chinese-Oriented Models			
Qwen3-Instruct	0.3766	0.0167	-95.57%
Doubao-1.5-Pro	0.6170	0.0216	-96.50%
GLM-4.5	0.4043	0.0085	-97.90%
DeepSeek-R1	0.6092	0.0621	-89.81%
Kimi-K2	0.6227	0.0043	-99.31%

The comparative results in Table 8 yield the following critical insights:

- **The Open-Ended Collapse:** We observe a catastrophic performance degradation from Phase 1-a (taxonomy-guided) to Phase 1-b (open-ended) across almost all tested models. Excluding Gemini-3, the average accuracy drop exceeds 90%. This suggests that while LLMs can effectively utilize a provided taxonomy as a reference for recognition, they lack the internal parameterized knowledge required to generate precise, hierarchical legal terms from scratch. In the open-ended setting, models frequently hallucinate nonexistent causes of action or fail to conform to the standard 282-*Cause of Action* nomenclature.

- **Gemini-3's Outlier Performance:** Gemini-3 Pro exhibits a significant advantage in the open-ended setting, achieving 0.3430 accuracy—nearly an order of magnitude higher than its peers. This indicates a superior alignment with Chinese legal taxonomies during its pre-training or instruction-tuning phase, enabling it to maintain structural fidelity without external taxonomy guidance.

- **Recall vs. Recognition:** Most Chinese-oriented models (e.g., Doubao-1.5-Pro, Kimi-K2) show strong "recognition" capabilities in Phase 1-a, matching the performance of top-tier global models like Llama 3.3 70B. However, their "recall" in Phase 1-b remains fragile. This confirms that localized pre-training primarily enhances the ability to process and select from domain-specific context rather than solving the "term recall" bottleneck in isolation.

F.2 One-Shot Performance Analysis

Overall, we observe that one-shot prompting exerts a relatively marginal influence across the majority of sub-benchmarks, with performance deltas frequently remaining near zero. To further investigate the impact of In-Context Learning (ICL) on the JurisBench pipeline, we conducted one-shot experiments on a randomly sampled subset of 100 cases. For each sub-benchmark, a single representative legal case with its corresponding "gold" reasoning and results was provided in the prompt as a demonstration.

As illustrated in Table 9, while the overall shifts are limited, the introduction of a single demonstration yields localized but significant gains for certain models. Notably, multilingual models like Llama 3.3 70B and Grok 4 exhibit substantial improvements in structural alignment (Phase 1) and judgment result determination (Phase 4-2), suggesting that ICL effectively helps these models adapt to the specific formatting requirements of Chinese judicial documents. However, we also observe a "sensitivity trap" in certain Chinese-oriented models, such as Qwen3-Instruct, which showed a sharp decline in Phase 3-4. This suggests that while one-shot prompting can clarify task constraints, it may also introduce biased priors that interfere with the internal legal reasoning of models already heavily aligned with domestic legal corpora.

F.3 Experiment of Law-specific Models

Preliminary Evaluation on Domain-Specific LLMs. We conducted a pilot study on several representative legal-specific models, including DISC-LawLLM (Yue et al., 2023), AdaptLLM/Law-Chat (Cheng et al., 2023), and Fuzi-Mingcha (Deng et al., 2023), using a subset of 100 cases of the Simple difficulty subset. The results revealed significant challenges in these models' stability across our multi-phase tasks. While they could partially complete basic tasks (e.g., Phase 1, 2, and 4-3), more than 50% of the outputs in remaining tasks exhibited severe **unfaithful generation**. Specifically, we observed frequent **instruction following failures** and symptoms of **model drift**, where models optimized for specific legal corpora failed to adapt to the structured, multi-step reasoning required by our benchmark.

Furthermore, these models often suffered from **textual degeneration** and semantic collapse. For instance, in Phase 3-2 (Legal Grounds-Law), Fuzi-Mingcha frequently failed to distinguish between legal entities and case evidence. Instead of outputting standardized legal categories, it generated a disorganized list of trial-related artifacts such as "xxx Hospital Discharge Records" and "Inpatient Bills" within the predicted legal name fields. Due to this high rate of unreliable outputs and the inability to maintain logical consistency, we opted to exclude these models from the full-scale large-scale evaluation to ensure the integrity of our comparative analysis, focusing instead on models that demonstrate sufficient functional stability.

Table 9: Experimental results (one-shot) of 10 LLMs.

		P1	P2	Phase 3: *Rationale of the Judgment*				Phase 4: *Result of the Judgment*		
Group	**Model**	1	2	3-1	3-2	3-3	3-4	4-1	4-2	4-3
Multilingual	Gemini-3 Pro	+0.000	+0.085	-0.006	+0.007	+0.036	+0.021	-0.002	+0.047	+0.124
	Claude 4.5	+0.209	+0.060	+0.025	+0.043	+0.101	+0.006	+0.016	+0.011	+0.014
	GPT-4o-mini	+0.010	-0.009	+0.006	+0.007	+0.003	+0.017	+0.005	+0.034	+0.006
	Grok 4	+0.267	+0.244	-0.019	+0.188	-0.114	+0.036	-0.026	+0.254	+0.030
	Llama 3.3 70B	+0.400	-0.031	-0.018	+0.176	+0.023	+0.035	-0.047	+0.446	-0.104
Chinese-Oriented	Qwen3-Instruct	+0.000	+0.070	-0.100	+0.373	+0.027	-0.593	+0.183	+0.067	+0.050
	Doubao-1.5-Pro	-0.031	+0.046	+0.005	+0.175	+0.013	+0.049	-0.021	+0.255	+0.124
	GLM-4.5	+0.188	-0.027	-0.003	+0.045	+0.017	+0.029	+0.025	+0.204	+0.094
	DeepSeek-R1	+0.030	+0.082	-0.025	+0.004	-0.296	-0.011	+0.044	+0.055	+0.064
	Kimi-K2	-0.099	+0.016	-0.035	+0.003	-0.342	-0.011	+0.005	+0.017	+0.093

F.4 Keyword-based Evaluation of Generative Tasks

To further validate the substantive legal precision of model outputs beyond linguistic fluency, we introduced a **Keyword-based Lexical Accuracy** metric for the four generative phases: *Focus of Disputes* (Phase 2), *Factual Basis* (Phase 3-1), *Rationale of the Judgment* (Phase 3-4), and *Result of the Judgment* (Phase 4-2).

Comparative Results. Table 10 presents the comparison between the original semantic metrics (reported in the main text) and the keyword-based lexical metrics.

Key Observations. The comparison reveals a sharp **Fluency-Precision Gap**: while models achieve high semantic scores (> 0.9), keyword recovery collapses in complex reasoning phases (≈ 0.2). Notably, Chinese-oriented models consistently outperform their multilingual counterparts in keyword recall, demonstrating higher *terminological fidelity*. While multilingual models often maintain semantic correctness, they tend to employ general descriptions rather than the precise Chinese legal nomenclature required for professional adjudication. Qwen3-Instruct's consistent lead highlights the efficacy of localized pre-training for professional grounding.

F.5 Extended Evaluation Across Case Complexity Tiers

To further investigate whether human-defined judicial difficulty corresponds to model performance degradation, we conducted extended experiments using Qwen3-Instruct and Gemini-3 Pro on the **Intermediate** and **Complex** subsets of JurisBench. The results (Table 11) reveal a surprising consistency: the models do not exhibit a significant performance drop as case complexity increases. In fact, *Qwen3-Instruct* achieves a Product metric of 82.6‰ on Complex cases, compared to 49.5‰ on Simple cases, while *Gemini-3* maintains performance within a similar single-digit percentage range across all tiers.

We provide the following analysis to explain this lack of variance between complexity tiers:

- **Divergence of Cognitive Load vs. Computational Constraints:** For human judges, complexity is a function of *workload*—higher difficulty typically involves more parties (multi-party litigation), voluminous evidence, and intricate factual causality, all of which increase human cognitive fatigue. Conversely, for LLMs, these factors primarily increase input token length. Given the extensive context windows of modern models, the "volume" of a case does not pose a fundamental challenge as long as the critical information remains within the context.

- **The Invariant Nature of Legal Logic Bottlenecks:** The primary bottlenecks identified in JurisBench—specifically *Statutory Symbolic Grounding* (Phase 3-3) and *Reasoning Stability*—are intrinsic to the legal task itself rather than the scale of the case. Whether a motor vehicle accident involves two parties or ten, the underlying requirement to map specific factual injuries to precise articles of the Civil Code remains equally rigorous. The model hits a "competence ceiling" at these logic-intensive nodes, which act as a universal filter regardless of the case's human-defined difficulty.

- **Paradox of "Complex" Case Performance:** The slight performance increase observed in

Table 10: Comparison between Original Semantic Metrics (Main) and Keyword-based Metrics (KW) across generative phases. "Main" refers to Embedding-based Similarity & LLM-as-a-Judge scores, while "KW" indicates the keyword recovery rate. All scores are normalized to $[0, 1]$. Values in **bold** and underline indicate the best and second-best performance in each **KW** column.

		Phase 2 (Focus)		Phase 3-1 (Facts)		Phase 3-4 (Reasoning)		Phase 4-2 (Result)	
Group	**Model**	Main	KW	Main	KW	Main	KW	Main	KW
Multilling.	Gemini-3 Pro	0.684	0.459	0.917	0.558	0.919	0.202	0.693	**0.269**
	Claude 4.5	0.644	0.401	0.901	0.455	0.938	0.155	0.446	0.159
	GPT-4o-mini	0.624	0.419	0.897	0.502	0.909	0.174	0.593	0.209
	Grok 4	0.681	0.521	0.901	0.424	0.926	0.088	0.711	0.192
	Llama 3.3 70B	0.782	0.559	0.917	0.540	0.923	0.165	0.604	0.151
Chinese-Oriented	Qwen3-Instruct	**0.798**	**0.656**	0.892	**0.639**	0.924	**0.286**	0.710	0.225
	Doubao-1.5-Pro	0.679	0.554	0.909	0.620	0.909	0.178	0.700	0.195
	GLM-4.5	0.738	0.577	0.912	0.556	0.922	0.223	0.776	0.209
	DeepSeek-R1	0.683	0.507	**0.927**	0.596	0.927	0.216	0.480	0.191
	Kimi-K2	0.750	0.571	0.904	0.493	**0.930**	0.205	**0.807**	0.219

Table 11: Model Performance on Intermediate and Complex Tiers.

		P1	P2	Phase 3				Phase 4			
Tier	**Model**	1	2	3-1	3-2	3-3	3-4	4-1	4-2	4-3	**Product**
Inter.	Qwen3-Instruct	0.987	0.779	0.925	0.667	0.383	0.925	0.735	0.691	0.878	75.0‰
	Gemini-3 Pro	0.865	0.708	0.928	0.691	0.564	0.905	0.761	0.742	0.880	99.7‰
Comp.	Qwen3-Instruct	0.967	0.779	0.925	0.669	0.380	0.925	0.790	0.677	0.943	82.6‰
	Gemini-3 Pro	0.962	0.673	0.943	0.724	0.223	0.918	0.824	0.738	0.808	44.5‰

some complex cases (e.g., Qwen3's performance in Phase 4-3) may be attributed to the richer context provided in complex judicial documents. More detailed evidence and rationales in the source text might, counter-intuitively, provide more "semantic anchors" for the model to leverage, whereas simple cases with sparse narratives may offer fewer cues for deep reasoning.

These findings underscore a critical insight for Legal AI: traditional case-weighting systems used for human judges are insufficient for evaluating LLMs. The benchmarks for AI must focus on the **logical depth of procedural transitions** rather than the mere volume of factual information.

G Human Subjects, Annotation Process, and Compensation

This study involves human participation in dataset construction and model evaluation. All procedures were designed to ensure objectivity, role separation, and auditability, while minimizing subjective bias.

G.1 Compensation and Labor Remuneration

All participants received fixed compensation independent of task outcomes or model performance. Student annotators were compensated at a rate of 100 RMB per person, reflecting standard research assistance rates. Judicial professionals were compensated at 200 RMB per person, reflecting their professional expertise and time commitment. No performance-based incentives were provided, and compensation was not contingent on specific annotation content or evaluation results.

G.2 Dataset Construction (Student Annotators)

A total of 9 student annotators (demographics in Table 12) were divided into two functionally independent groups to ensure data integrity through a "construction-verification" pipeline.

Roles and Process. Six students formed the **Construction Group**, responsible for anonymizing personal identifiers, structuring raw judgments into the JurisBench schema, and screening outcome-relevant keywords (e.g., liability attribution, causal relations). Three students formed the **Verification Group**, who independently reviewed the structured outputs for consistency and completeness. Discrepancies were resolved through predefined consistency rules via a double-blind process.

Table 12: Demographic Information of Student Annotators

Student ID	Age	Gender	Work Location
Student 1	23	Male	Beijing
Student 2	23	Male	Beijing
Student 3	25	Female	Beijing
Student 4	28	Male	Beijing
Student 5	28	Male	Beijing
Student 6	28	Female	Beijing
Student 7	35	Female	Beijing
Student 8	29	Male	Beijing
Student 9	24	Male	Beijing

Guidelines. Annotators were strictly prohibited from introducing interpretative judgments or reconstructed reasoning. Structured fields were required to reflect explicit content from the source judgment. Anonymization was limited to removing personal identifiers without altering legally relevant facts.

G.3 Human Evaluation and Validation (Judicial Professionals)

Fifteen judicial professionals (demographics in Table 13) participated in evaluating model performance and validating our automated metrics.

Table 13: Demographic Information of Judicial Professionals

Judge ID	Age	Gender	Work Location
Judge 1	42	Female	Beijing
Judge 2	36	Female	Beijing
Judge 3	33	Female	Beijing
Judge 4	36	Female	Beijing
Judge 5	45	Female	Beijing
Judge 6	45	Male	Beijing
Judge 7	27	Male	Shandong
Judge 8	37	Female	Shandong
Judge 9	32	Female	Shandong
Judge 10	27	Female	Shandong
Judge 11	36	Female	Shandong
Judge 12	28	Male	Shandong
Judge 13	29	Male	Shandong
Judge 14	34	Male	Shandong
Judge 15	28	Male	Shandong

Evaluation Setting. Judicial professionals acted as judicial assistants under a strict blind-review protocol. They were not informed whether an answer was produced by a human or an LLM. All answers were presented in a uniform format to eliminate stylistic or metadata cues. Evaluators assessed model predictions for the *Focus of Disputes* task under two conditions: with and without access to the "gold" reference extracted from the original document.

Judicial Evaluation Protocol and Human-Metric Alignment. To validate our automated *LLM-as-a-Judge* framework (which provides continuous scores in $[0, 1]$), we conducted a human-in-the-loop study on Phase 2 (*Focus of Disputes*). We adopted a discrete seven-point ordinal scale for judges to better mirror practical judicial decision-making: **0**: Completely incorrect; **1**: Extensive errors, unusable; **2**: Substantial errors, no meaningful assistance; **3**: Approximately half correct; **4**: Mostly correct, requiring minor corrections; **5**: Sufficiently accurate for adjudication; **5+**: Functionally equivalent to high-quality professional output.

Consistency and Validity. To ensure objectivity, we selected the answers of the tested LLMs of five representative cases for a cross-validation study, where each case was evaluated by five independent judges (25 total assessments). Evaluators focused on substantive legal alignment rather than surface-level fluency. We then mapped the continuous LLM ratings onto this ordinal scale and performed a concordance analysis. The results demonstrate high consistency between the judges' consensus and the automated scores, confirming that our framework effectively serves as a reliable proxy for professional-level judicial reasoning. Judges treated the reference *Focus of Disputes* as the normative baseline and assessed each case in isolation to prevent cross-referencing bias.

H Written Guidelines for Student Annotators

This section provides the complete and operational written guidelines for student annotators. The objective of this task is to transform raw judicial documents into a high-fidelity and structured dataset through two primary processes: **Data Structuring** and **Outcome-Relevant Keyword Labeling**.

H.1 Role Definition and Workflow

Student annotators act as technical data extractors. Your objective is to perform a high-fidelity mapping of the judicial text into a structured schema.

- **Verbatim Extraction:** You must use the exact wording from the judgment. You are strictly prohibited from performing any paraphrasing or summarizing.
- **No Linguistic Cleaning:** If the phrasing of the judge is awkward or grammatically imperfect, you must preserve it exactly as it appears.
- **Mental Firewall:** You must disregard your personal knowledge of the law. If a fact is missing in the text, it must be recorded as *Not Specified*.

H.2 Task I: Verbatim Data Structuring

Annotators must decompose the judgment into seven specific fields using a **Key-Value pair** format. The *Key* represents the attribute of the information segment, and the *Value* is the verbatim text extracted from the document.

The Seven Mandatory Fields:

1. ***Basic Case Info* Information:** This includes metadata such as the Case Name, the *Cause of Action*, the *Court of Acceptance*, the *Parties* and their Procedural Roles, the Procedural Posture, and the Date of Case Closure.
2. ***Plaintiff's Arguments*:** This includes the claims, the factual grounds, and the litigation requests explicitly raised by the plaintiff or plaintiffs.
3. ***Defendant's Arguments*:** This includes the responses, the defenses, the objections, or the admissions explicitly raised by the defendant or defendants.
4. ***Facts Proved by the Court*:** This consists of the objective facts and evidence determined and verified by the court.
5. ***Focus of Disputes*:** This identifies the core points of contention or the disputed issues as defined by the court.
6. ***Rationale of the Judgment*:** This covers the reasoned analysis of the court, the application of law to the established facts, and the legal logic.
7. ***Result of the Judgment*:** This is the final adjudicative outcome and the specific orders of the court.

Structuring Rules:

- **Format Requirement:** Every entry must be represented as: Attribute_Name: "Exact text from the document".
- **Strict Copying:** You must not change a single character. If a segment of text does not logically fit into any structured attribute, you must abandon the structuring for that specific segment rather than forcedly adapting the content.
- **Missing Information:** If a field is not explicitly stated in the judgment, it must be recorded as *Not Specified*.

H.3 Task II: Outcome-Relevant Keyword Labeling

Keywords must be labeled within three specific fields: ***Focus of Disputes***, ***Rationale of the Judgment***, and ***Result of the Judgment***. Annotators must identify all information that could materially influence the adjudicative outcome of the case.

Labeling Categories and Examples:

Annotators must identify and label the following categories:

- **Legal Terms:** Doctrinal concepts and terminology, for example, *tort* or *liability*.
- **Legal Statutes and Articles:** Specific legal citations, for example, *Civil Code Article 1165*.
- **Factual Elements:**
  - **Verbs:** Actions describing the core events, for example, *crash* or *breach*.
  - **Nouns:** Key objects, entities, or locations, for example, *hospital* or *contract*.
  - **Amounts:** Specific monetary values, percentages, or numbers, for example, *50,000 Renminbi* or *10 percent*.
- **Involved Personnel:** This includes all parties involved in the litigation and identified individuals relevant to the facts.

Negative Constraints: To ensure the purity of the data, do not label words that represent personal writing habits or logical transitions that do not hold substantive legal weight, such as *through*, *therefore*, *accordingly*, or *moreover*.

I One-Page Annotation Checklist

```
        ONE-PAGE ANNOTATION CHECKLIST

[PHASE ONE: STRUCTURING AUDIT]
- Is the data organized into the seven
    required fields?
- Is every entry formatted as a Key-Value
    pair?
- Does the extracted text match the
    original document character
  for character?
- Have you avoided any paraphrasing or
    summarizing?
- Have you cleaned the text of digital
    artifacts (e.g., page
  numbers, headers)?

[PHASE TWO: KEYWORD AUDIT]
- Have keywords been labeled only in:
  * Focus of Disputes
  * Rationale of the Judgment
  * Result of the Judgment
- Have you included all parties involved
    and identified personnel?
- Have you included all monetary amounts
    and specific legal articles?
- Have you excluded stylistic transition
    words (e.g., therefore)?

[FIELD-SPECIFIC REFERENCE GUIDE]

# Base Information
Extract from the document header and
    initial paragraph.
(No keyword labeling required)

# Plaintiff's Arguments
Extract from the section starting with the
    claims of the plaintiff.
(No keyword labeling required)

# Defendant's Arguments
Extract from the section starting with the
    defenses of the defendant.
(No keyword labeling required)

# Facts Proved by the Court
Extract from the section describing the
    established facts.
(No keyword labeling required)

# Focus of Disputes
Extract from the judicial summary of
    disputed issues.
Labeling focus: legal terms, disputed
    actions, specific amounts.

# Rationale of the Judgment
Extract from the section containing the
    opinion of the court.
Labeling focus: legal statutes, legal
    doctrines, key nouns.

# Result of the Judgment
Extract from the final adjudicative orders.
Labeling focus: final amounts, party names,
    final outcomes.
```

J Verification Principles and Review Procedure

Role of Verification Annotators:

Verification annotators act as rule-based auditors rather than secondary annotators. Prior to review, they must study and understand all annotation guidelines and treat them as the sole normative standard.

Scope of Verification:

Verification requires checking:

- Completeness of required fields
- Consistency across structured sections
- Fidelity to the original judicial text
- Compliance with all annotation prohibitions

Verification annotators must not optimize wording or introduce legal judgment.

Error Identification Criteria:

An annotation is considered problematic if required information is missing, unsupported by the source text, improperly inferred, or placed in an incorrect field.

Error Logging and Correction Protocol:

When an issue is identified, verification annotators must record:

- Case ID
- Affected Field
- Problem Description
- Rule Violated
- Revised Content

Standard Error Log Template:

- Case ID:
- Affected Field:
- Problem Description:
- Rule Violated:
- Revised Content:

All corrections must strictly follow the original annotation rules and remain fully traceable to the source text.

Resolution Principle:

Corrections must be made strictly according to written rules. When multiple revisions are possible, the version that minimizes interpretation and maximizes textual fidelity must be selected.

Auditability Statement:

The structured error logs ensure that all annotation revisions are auditable, transparent, and reproducible, enabling independent assessment of dataset construction quality.