To Judge or Not to Judge: Can Large Language Models Leverage the Dispute Focus in Legal Judgment?

Luoming Hu[1], **Liang Yang**[2,3,†], **Jingjie Zeng**[2], **Zijie Xing**[1]
[1]School of Future Technology, Dalian University of Technology, China
[2]School of Computer Science and Technology, Dalian University of Technology, China
[3]Key Laboratory of Social Computing and Cognitive Intelligence, Ministry of Education, China
{huluoming2004, jjtail, kokojjj}@mail.dlut.edu.cn, liang@dlut.edu.cn

Abstract

Civil judicial cases are highly complicated, posing significant challenges for Large Language Models (LLMs) for Legal Judgment Prediction (LJP). While judges manage this complexity through the dispute focus—a mechanism distilling cases into core issues—existing research largely overlooks this tool in favor of generic reasoning frameworks that lack authentic judicial logic. To bridge this gap, we first introduce **FocalLaw**, the first dataset aligning full-process Chinese civil judicial data through the dispute focus, comprising 1,000 high-quality cases across six causes of action. Building on this dataset, we examine LLMs' capability to utilize the dispute focus and uncover a counter-intuitive phenomenon: LLMs fail to leverage the dispute focus even with CoT and SFT, which we identify as the "Clerk Trap". To solve the problem, we propose **FocalJudge**, a novel framework that leverages the dispute focus to guide LLMs through a structured, judge-like cognitive workflow. Experimental results demonstrate the effectiveness of FocalJudge and offer valuable insights into the interpretability and reliability of LLMs in the legal domain.

1 Introduction

In recent years, LLMs have increasingly been deployed to augment legal practice (Dahl et al., 2024; Wang et al., 2025), yet the demand for their reliability in complex adjudication tasks remains largely unmet (Ariai and Demartini, 2025; Shao et al., 2025). Civil judicial cases, characterized by lengthy documents, diverse evidence, and intricate legal applicability, pose significant and unique challenges to existing LLMs in LJP. Consequently, there is an urgent need for methods that can make effective reasoning and judgments on complex civil judicial cases using appropriate legal provisions.

To navigate such complexity and enhance interpretability, various efforts have been made, including interpretable Chain-of-Thought (CoT) prompts (Fujita et al., 2024), syllogistic prompting techniques (Deng et al., 2023) and neuro-symbolic systems (Wei et al., 2025). However, these methods tend to impose generic reasoning frameworks derived from general domains, thereby neglecting authentic judicial proceedings. Specifically, they overlook the dispute focus—the core cognitive mechanism judges use to organize a chaotic case into several core issues.

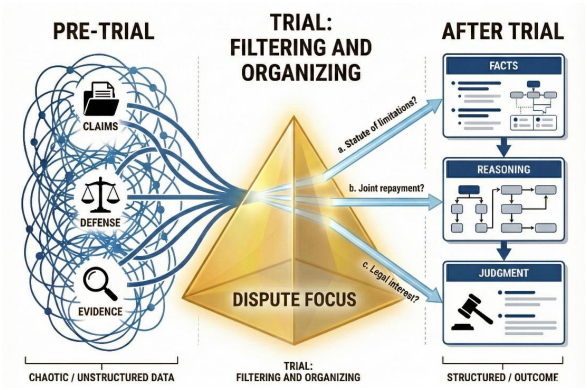

Figure 1: The Dispute Focus acts as a cognitive "dimensionality reduction" mechanism, transforming chaotic pre-trial data into structured judicial outcomes.

We believe that the dispute focus is not just a procedural tool, but a cognitive "dimensionality reduction" mechanism. It transforms complex, multimodal, redundant, and adversarial judicial proceedings into a set of clear, verifiable propositions, as illustrated in Figure 1. For LLMs, the ability to effectively utilize this mechanism is a key test of whether they can truly understand and engage in complex legal reasoning, rather than simply performing superficial text processing.

To validate the effectiveness of the dispute focus, a dataset that covers the entire judicial proceeding—from initial pleadings and evidence, through the dynamic court debate process, to the final written legal case document, and is centrally annotated with the dispute focus—is indispensable. Existing

†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30550–30571

Dataset	Case Type	Pre-Trial	Court Debate	After-Trial	Dispute Focus	#JC	#Cases	Avg. Length
CAIL2018 (Xiao et al., 2018)	Criminal	✓				✓	1.5M+	329
MSJudge (Ma et al., 2021)	Civil	✓	✓			✓	70,482	1k
LegalBench (Guha et al., 2023)	Civil	✓		✓			91,206	486
CaseHOLD (Chalkidis et al., 2022)	Both			✓		✓	52,800	245
CFDS (Duan et al., 2019)	Civil		✓		✓		5,477	-
FocalLaw (Ours)	Civil	✓	✓	✓	✓	✓	1,000	**19,379**

Table 1: Comparison of **FocalLaw** with existing legal datasets. Columns indicate the inclusion of *Pre-Trial* information (e.g., Complaint, Defense), *Court Debate* transcripts, *After-Trial* documents (e.g., Judgment, Reasoning), and the *Dispute Focus* annotation. **#JC** denotes the Judgment Class, **#Cases** denotes the number of samples, and **Avg. Length** represents the text length scale.

legal datasets, while valuable, are insufficient for modeling the complete judicial reasoning chain. As demonstrated in Table 1, they either lack crucial procedural stages (e.g., pre-trial documents, after trial results) or contain no annotations for the dispute focus, preventing models from learning how core controversies translate into final judgments. There is still a lack of a comprehensive, multi-stage legal dataset that fits our needs.

To bridge this gap, we construct FocalLaw, the first dataset to align full-process Chinese civil judicial data through the dispute focus. This dataset comprises 1,000 high-quality civil cases across six common causes of action, meticulously aligned from 730,000 judgment-video pairs through a rigorous four-stage filtering process, yielding a retention rate $< 0.14\%$. Critically, the dispute focus is directly extracted from legal documents where it was generated by presiding judges and both parties, ensuring authenticity and procedural fidelity. To model the authentic, structured procedure of judicial proceedings centered on the dispute focus, we design a series of research tasks based on the dataset to answer the following core questions:

RQ1: How do LLMs respond when they are informed of the dispute focus?

RQ2: Can LLMs leverage dispute focus when determining facts and making judgments?

RQ3: How to make a better use of the dispute focus?

Through our experiment in RQ2, we uncover the "Clerk Trap", where informing LLMs of the dispute focus degrades performance by encouraging summarization over adjudication. Specifically, CoT exacerbates this trap by overly reinforcing the dispute focus without human guidance, while SFT fails to bridge this reasoning gap due to the extensive length and high heterogeneity of civil cases. To address these problems, we propose FocalJudge, a novel framework that leverages the dispute focus to guide LLMs through a structured, judge-like cognitive workflow. The main contributions of this paper are summarized as follows:

- We construct the first dataset to align full-process Chinese civil judicial data through the dispute focus, containing 1000 high-quality cases, which can be used to systematically analyze the role of the dispute focus in judgment prediction and the entire judicial proceedings.
- We uncover the "Clerk Trap" phenomenon, where LLMs excel at summarizing conflicting claims like a "clerk" but fail to adjudicate evidence and determine facts like a "judge". Furthermore, we demonstrate that standard optimization techniques, such as SFT, are insufficient to bridge this reasoning gap due to the complexity of civil cases.
- We propose a new framework, FocalJudge, which can effectively utilize the dispute focus for judgment prediction, providing new insights and solutions for applying artificial intelligence technology to assist in judicial adjudication in more realistic and complex scenarios.

2 Related Work

Legal Datasets Existing legal NLP research benefited from comprehensive benchmarks like LexGLUE, LegalBench, CaseGen, JuDGE (Chalkidis et al., 2022; Guha et al., 2023; Li et al., 2025; Su et al., 2025) and large-scale datasets CAIL2018 and MSJudge (Xiao et al., 2018; Ma et al., 2021; Duan et al., 2019).

While these datasets advanced the field, they universally lack the crucial element of the judge-summarized dispute focus that links the entire judicial proceeding. This prevented models from learning the complete reasoning chain from core controversy to final judgment, a gap that FocalLaw is specifically designed to fill.

LLM Reasoning in Legal Domain To navigate the inherent complexity of civil adjudication, various efforts have been made, ranging from CoT prompting (Fujita et al., 2024), syllogistic reasoning (Deng et al., 2023), neuro-symbolic framework (Wei et al., 2025), influence functions (Valvoda and Cotterell, 2024), rule-based legal systems (Billi et al., 2023) and evolutionary game theory (Pereira et al., 2024). However, these methods rely on generic logic rather than tapping into the inherent logic of the legal process itself, which is the approach our work pioneers.

3 Dataset Construction

The authentic judicial proceeding centers on the dispute focus: **a set of core questions distilled from the plaintiff's claims and the defendant's rebuttals by the judge and the two parties during the Pre-Trial stage.** It not only clarifies the core issues and organizes evidence, but also condenses claims and promotes settlement, serving as the logical hub that connects the claims to the final judgment. Thus, mastering the dispute focus is crucial for LLMs to perform true legal reasoning.

However, while the raw data for such analysis (legal documents and court debate) exist in the public domain, they are disparate, unstructured, and unaligned. To address this, we conduct the curation of a high-quality dataset that aligns full-process Chinese civil judicial data through the dispute focus, transforming these raw materials into a computable, research-ready format. A dataset sample is provided in Appendix A.

3.1 Data Source and Alignment

The foundation of FocalLaw is built upon two authoritative public sources: China Judgments Online [*] (which provides legal case documents) and China Court Trial Online [†] (which provides court debate videos). The primary challenge is that these sources are not directly linked. Our first contribution is to meticulously align these disparate sources, identifying true judgment-video pairs. We perform a rigorous four-stage filtering process:

In the first step, we collect approximately 730,000 judgment-video pairs. After the review process that removes duplicate case codes and checks the core judicial elements, only about 6,000 pairs remain. We then analyze the distribution of causes of action across all qualified cases and select the six most frequent categories: private lending disputes, sales contract disputes, motor vehicle accident liability disputes, labor disputes, contract disputes, and construction contract disputes. Together, these six categories cover 42.1% of the total sample, while none of the remaining 271 causes of action individually exceeds a 3% share. After extracting the six most common causes of action, 2,185 pairs are left. Finally, we remove those with poor audio quality and only 1000 pairs remain.

This process highlights the rarity of complete, high-quality, and matched case data, underscoring the value of the curated collection.

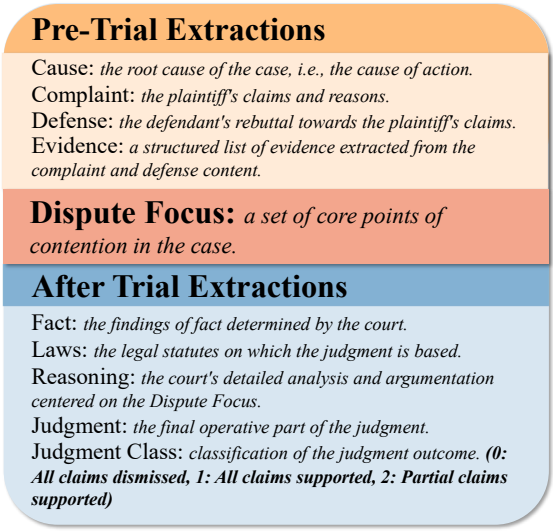

Figure 2: Detailed explanation of the meaning of each extracted field.

3.2 Data Annotation

3.2.1 Audio Transcription

To convert the court debate recordings into text that can be processed by LLMs, we use Whisper (Radford et al., 2022) to transcribe the qualified video recordings into text labeled with different speakers. To ensure the quality of the transcription, we randomly select 200 audio clips of trial recordings from the dataset (each clip lasting approximately 20 minutes) for manual transcription and evaluation. After evaluation, the model's Character Error Rate (CER) is found to be below 7.4%, which is sufficient to meet the needs of this research for analyzing the content of trial dialogues.

3.2.2 Structuring Legal Case Documents

The original legal documents are long-form, semi-structured texts, making them difficult to use for targeted computational analysis. Our main con-

[*] https://wenshu.court.gov.cn/
[†] https://tingshen.court.gov.cn/

All dismissed	All supported	Partial supported
0.82	0.62	0.69

Table 2: Fleiss' Kappa for different Classes.

tribution lies in transforming this raw text into a highly structured and verified format. We design a schema with 10 core fields to represent the entire judicial process: **Cause, Complaint, Defense, Evidence, Dispute Focus, Fact, Laws, Reasoning, Judgment and Judgment Class.** The detailed description of each field is shown in Figure 2. The population of this schema is a two-step, human-in-the-loop process: First, we utilize Qwen2.5-72b (Qwen, 2024; Yang et al., 2024) to perform an initial extraction and map the unstructured text into our predefined schema by designing specific prompts (see Appendix B). Second, recognizing the risk of hallucinations when LLMs process legal texts, we establish a strict manual review and correction stage to ensure the absolute accuracy of the data. All fields initially annotated by the LLM are individually checked and corrected by annotators with a legal background, with particular attention to the accuracy of key factual information such as Fact and Evidence. We also anonymize the personal information in the data during this process.

3.2.3 Annotation Quality Assurance

Our annotation team consists of three experts with legal background who underwent systematic training. To address the inherent subjectivity of the Judgment Class field, we implement a cross-annotation process where all three annotators independently label each document. The resulting inter-annotator agreement (IAA), measured using Fleiss' Kappa (Fleiss, 1971), is detailed by classification level in Table 2. Crucially, the Dispute Focus field—the most critical component of this dataset—is not generated by an LLM or our annotation team. Instead, it is discussed and recorded in the legal case document by the presiding judge and the two parties before the court debate, providing a robust foundation for our research. Furthermore, to prevent target leakage, we delete the dispute focus mentioned in Court Debate, Fact and Reasoning.

3.3 Dataset Statistics and Analysis

An analysis of the dataset's dispute focus, detailed in Table 3, reveals two primary characteristics. First, they are few in number, with an average of approximately two per case and a maximum of four. Second, they are remarkably concise, with the longest average length being only about 30 characters. This brevity reflects their function in helping judges distill core issues, organize evidence, and consolidate claims.

Statistic	#Count	#Length
Avg. Total	1.96	24.84
Avg. Private Lending	1.84	23.92
Avg. Sales Contract	1.86	26.46
Avg. Motor Vehicle Accident	1.88	20.33
Avg. Labor	2.07	26.78
Avg. Contract	2.06	31.74
Avg. Construction Contract	2.32	24.26

Table 3: Distribution of dispute focus in FocalLaw. #Count stands for the average count of dispute focus; #Length stands for the text length of the dispute focus.

Furthermore, the Judgment Class field exhibits a significant and expected class imbalance, with a distribution of approximately 2:1:7 for dismissed, fully supported, and partially supported, respectively. In complex cases, courts rarely grant a total victory or a complete dismissal. Instead, they typically validate some claims while rejecting others, making partial support (Class 2) the most common result. This pattern aligns with findings from other legal datasets, such as the European Court of Human Rights (ECHR), where absolute wins or losses are a minority of judgments (Trautmann et al., 2022). From a model evaluation perspective, this class imbalance provides us with invaluable hard cases. Although samples from Class 0 and 1 are less frequent, they are crucial to testing whether LLMs are genuinely leveraging the dispute focus.

4 Experiments

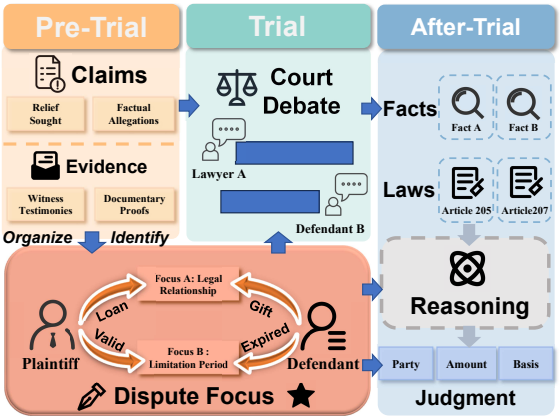

Figure 3: Overview of the alignment of the full judicial proceeding centered on the Dispute Focus.

Experiment Settings RQ1 investigates whether, when facts are undetermined, the dispute focus can serve as an attentional guide, helping the model to make preliminary inferences from the source material. RQ2 further breaks down judicial proceedings into two stages: fact determination and law application, exploring whether the dispute focus interferes with the model's ability to determine facts and apply law to established facts. For RQ1 and RQ2, we assess model performance by running experiments that either include or exclude the dispute focus from the input. The RQ3 then compares the performance of our proposed framework against a baseline that does not use the dispute focus.

Our selection of inputs for each RQ is designed to mirror the distinct stages of judicial decision-making, as illustrated in Figure 3. The Pre-Judgment input, which consists of raw case materials (Cause, Complaint, Defense, Evidence, and Court Debate) without any official conclusions, simulates the pre-judgment phase where a judge faces claims, evidence and the debate between the two parties to determine the fact. The Post-Fact input includes Cause, Complaint, Defense, Laws, and the established Fact, which simulates the subsequent stage where a judge applies relevant laws to determined facts to formulate a final judgment.

Based on the input design above, we set the Pre-Judgment input to be used in RQ1, Fact generation in RQ2 and RQ3. The Post-Fact input is used in Reasoning, Judgment, and LJP in RQ2 and RQ3.

To ensure the fairness and validity of the comparison results, we only make minimal modifications to the prompts for the experimental group compared to the control group. More details on prompt design can be found in Appendix C. For closed-source LLMs, we conduct experiments by calling their APIs. Inputs that exceed the maximum context of the LLM, we truncate from the middle to ensure the maximum retention of effective information because the beginning and the end may contain crucial information. The temperature of all LLMs is set to 0 to ensure reproducibility. We employ Low-Rank Adaptation (LoRA) in the SFT in RQ2 (Hu et al., 2021). More training details are provided in Appendix D.

Evaluation Metrics To evaluate the performance of LLMs, we employ a set of established metrics for both text generation and LJP. The text generation includes Fact, Reasoning and Judgment generation. The LJP includes a three-class classification task. For text generation evaluation, we use *ROUGE-L* (Lin, 2004), *BERTScore* (Zhang et al., 2019), and *LLM evaluation*. We use *GPT-4o* as the LLM judge to evaluate the performance of other LLMs (Li et al., 2025). For LJP evaluation, we report Macro F1, Weighted F1 and Accuracy. More details and validation can be found in Appendix E.

Baselines To comprehensively evaluate LLMs capability to leverage the dispute focus, we select 10 well-known models with varying parameter sizes and capabilities in three categories: **Closed-Source LLMs:** Claude-3.5-Sonnet (Anthropic, 2024), DeepSeek-v3 (Liu et al., 2024), Gemini-3-Pro (DeepMind, 2025), Grok-4 (xAI, 2025), GPT-4o (OpenAI, 2024); **Open-Source LLMs:** Gemma3-12B (Gemma, 2025), Llama-3.1-8B-Instruct (AI@Meta, 2024), Qwen3-8B (Yang et al., 2025); **Legal-Domain LLMs:** ChatLaw (Cui et al., 2024), LexiLaw (Li et al., 2024).

5 Results and Analysis

5.1 RQ1: How do LLMs respond when they are informed of the dispute focus?

To answer RQ1, we conduct a comparative experiment, aiming to evaluate the impact of dispute focus on generating Reasoning, Judgment, and LJP by either including or excluding it from the input. The input for this task intentionally mirrors the complete, unprocessed information a judge initially receives, which is the Pre-Judgment input.

Based on the experimental results in Table 4, informing LLMs of the dispute focus leads to a modest performance improvement across most LLMs in tasks including Reasoning, Judgment, and LJP. This trend is observed across different model tiers; top-tier LLMs like Gemini-3 and GPT-4o, smaller open-source models such as Gemma3, Llama-3.1-8B, and Qwen3-8B, and legal-domain specific models like ChatLaw and LexiLaw, all demonstrate generally enhanced performance with the inclusion of the dispute focus. Specifically, Gemini-3 achieves the highest scores in both Reasoning and LJP tasks. In the Judgment task, performance varies, with different models excelling in metrics like ROUGE-L, BERTScore, and LLM evaluation scores. Overall, the general trend indicates that adding the dispute focus as an input consistently, though limitedly, improves model performance across the board.

Model	Reasoning			Judgment			LJP		
	ROU.	BS.	LLM	ROU.	BS.	LLM	MF1.	WF1.	Acc.
Small LLMs (with Basic Prompt and 1 Example)									
Gemma3	17.50 $\downarrow_{0.34}$	70.84 $\uparrow_{0.52}$	3.84 $\uparrow_{0.12}$	23.41 $\uparrow_{0.55}$	75.49 $\uparrow_{2.5}$	3.08 $\uparrow_{0.05}$	43.03 $\uparrow_{4.11}$	67.05 $\uparrow_{7.09}$	60.00 $\uparrow_{6.84}$
Llama3.1-8B	18.47 $\uparrow_{0.53}$	70.94 $\uparrow_{1.34}$	3.66 $\uparrow_{0.02}$	26.36 $\uparrow_{0.42}$	72.12 $\downarrow_{0.96}$	2.99 $\uparrow_{0.11}$	37.04 $\downarrow_{0.86}$	59.27 $\downarrow_{0.81}$	52.76 $\downarrow_{1.55}$
Qwen3-8B	24.07 $\uparrow_{1.52}$	72.12 $\uparrow_{0.96}$	4.11 $\uparrow_{0.10}$	33.98 $\uparrow_{2.03}$	78.34 $\uparrow_{0.33}$	3.82 $\uparrow_{0.03}$	45.04 $\uparrow_{0.25}$	68.48 $\uparrow_{0.85}$	62.62 $\uparrow_{1.51}$
ChatLaw	12.71 $\uparrow_{0.34}$	66.13 $\uparrow_{0.67}$	2.67 $\uparrow_{0.13}$	17.62 $\uparrow_{0.24}$	68.17 $\downarrow_{0.14}$	2.92 $\uparrow_{0.04}$	30.43 $\uparrow_{3.50}$	53.09 $\uparrow_{0.82}$	51.25 $\uparrow_{1.67}$
LexiLaw	8.18 $\downarrow_{1.09}$	58.62 $\uparrow_{0.46}$	2.28 $\uparrow_{0.05}$	14.40 $\uparrow_{0.47}$	59.58 $\downarrow_{0.15}$	1.97 $\uparrow_{0.03}$	21.98 $\uparrow_{1.67}$	45.39 $\uparrow_{0.36}$	48.34 $\uparrow_{0.50}$
LLM APIs (with Basic Prompt and 1 Example)									
Claude-3.5	24.06 $\uparrow_{1.51}$	72.11 $\uparrow_{0.95}$	4.63 $\uparrow_{0.18}$	25.78 $\downarrow_{2.07}$	79.52 $\downarrow_{1.12}$	3.58 $\downarrow_{0.02}$	58.00 $\uparrow_{2.30}$	75.37 $\uparrow_{0.11}$	73.00 $\uparrow_{0.00}$
DeepSeek-v3	26.17 $\downarrow_{0.76}$	71.18 $\uparrow_{0.38}$	3.63 $\uparrow_{0.22}$	34.47 $\downarrow_{0.34}$	**80.19** $\uparrow_{1.19}$	3.55 $\uparrow_{0.03}$	57.27 $\uparrow_{1.39}$	78.59 $\uparrow_{1.75}$	77.89 $\uparrow_{1.01}$
Gemini-3	**32.29** $\uparrow_{1.90}$	**73.64** $\uparrow_{0.71}$	**5.29** $\uparrow_{0.04}$	33.63 $\uparrow_{0.93}$	80.00 $\uparrow_{0.12}$	**3.88** $\uparrow_{0.11}$	**67.48** $\uparrow_{2.79}$	**83.62** $\uparrow_{0.60}$	**83.42** $\uparrow_{0.50}$
Grok-4	23.64 $\uparrow_{1.56}$	72.05 $\uparrow_{1.03}$	4.47 $\uparrow_{0.16}$	28.78 $\uparrow_{0.78}$	78.04 $\downarrow_{0.49}$	3.61 $\uparrow_{0.03}$	47.31 $\uparrow_{3.50}$	71.52 $\uparrow_{1.83}$	69.59 $\uparrow_{1.55}$
GPT-4o	20.63 $\uparrow_{4.05}$	72.60 $\uparrow_{1.81}$	3.61 $\uparrow_{0.15}$	**35.81** $\uparrow_{1.24}$	78.64 $\uparrow_{0.56}$	3.75 $\uparrow_{0.23}$	61.37 $\uparrow_{3.51}$	80.60 $\uparrow_{1.86}$	79.70 $\uparrow_{1.64}$

Table 4: The results of RQ1. "ROU." represents the ROUGE-L score (%), "BS." stands for BERTScore (%), "LLM" refers to the scores assigned by the LLM Judge, "MF1." represents the Macro-F1, "WF1." represents the Weighted-F1 and "Acc." stands for Accuracy. "↑" signifies an improvement over the baseline, while "↓" indicates a decline. The best results are highlighted in bold.

5.2 Answer1: Consistent but limited improvement

Although directly informing LLMs of the dispute focus improves their performance, the effect is quite limited. We attribute this to two factors:

First, advanced models like GPT-4o and Gemini-3-Pro are already attempting to implicitly identify and locate the core conflicts between the parties embedded within complex legal texts. The dispute focus acts as an attention guide, helping the models confirm and refine their focus on the most critical issues rather than providing entirely new information. This explains why performance improves, but only to a limited extent.

Second, our LJP task in RQ1—classifying a judgment as "all claims supported," "partially supported," or "all claims dismissed"—requires more than just identifying the issues. The final judgment hinges on a detailed assessment of evidence for each dispute focus. The provided focus tells the model what questions to answer, but not how to answer them. The model must still independently perform the complex reasoning over facts and evidence, which explains the limited impact.

To conclude, this systematic analysis shows that informing LLMs of the dispute focus results in a consistent but limited improvement.

5.3 RQ2: Can LLMs leverage dispute focus when determining facts and making judgments?

For RQ2, we employ a two-stage comparative experiment with and without the dispute focus. For the Fact Generation task, we use the Pre-Judgment input. For the Reasoning, Judgment, and LJP tasks, we use the Post-Fact input. However, based on the experimental results, directly informing LLMs of the dispute focus leads to a counter-intuitive and unexpected decrease in performance across several tasks. The results are in Appendix F and I.

To further study the performance decline, we conduct a comparative experiment using Qwen3-8B and Gemma3, evaluating four strategies: Original (baseline), SFT (Supervised Fine-Tuning), CoT (Chain-of-Thought), and Dispute Focus (directly providing the dispute focus).

As shown in Figure 4, using CoT results in an even worse performance across all tasks, meaning that this decline is not resolved by specific prompting techniques. Furthermore, SFT also failed to solve the problem, indicating that the complexity within the civil cases cannot be simply solved by fine-tuning. More experimental details and case study can be found in Appendix G.

5.4 Answer2: The "Clerk Trap" and SFT Limitations in Civil Adjudication

The counter-intuitive decline in performance observed in RQ2, alongside the failure of standard optimization techniques, reveals two fundamental insights regarding the capabilities of current LLMs in civil judicial settings.

First, LLMs excel as **"clerks"** but fail to function as **"judges"** when directly informed of the dispute focus. We attribute the performance decline to a shift in the model's cognitive mode. When provided with the dispute focus, the LLM acts as a "clerk": it efficiently summarizes, induces, and re-

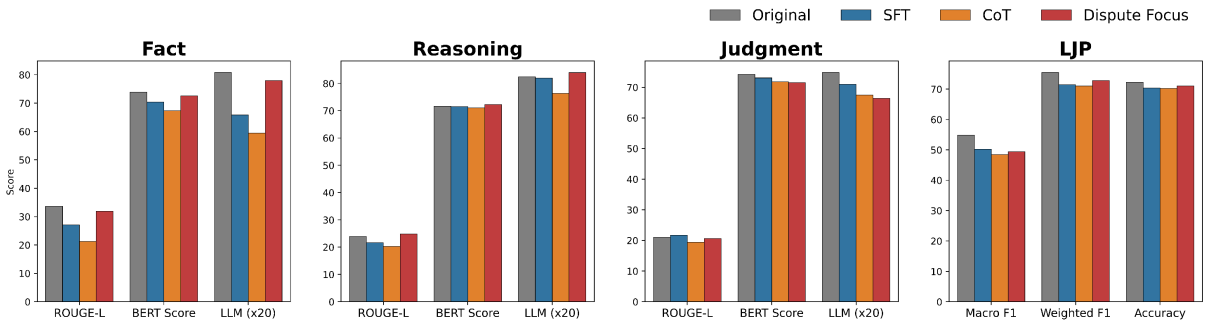

Figure 4: Comparative experimental results of four strategies across all tasks in RQ2. The results illustrate that standard optimization techniques like SFT and CoT, as well as directly providing the dispute focus, fail to consistently improve performance compared to the baseline, with CoT leading to a significant decline.

states the conflicting viewpoints of the plaintiff and defendant regarding the core issues. However, it fails to perform the role of a "judge": it lacks the capacity to weigh evidence, adjudicate between conflicting claims, and determine the objective truth. As illustrated by the case study in Appendix H, the model often confuses "unresolved disputes" with "established facts", indicating that without a structured cognitive workflow, the dispute focus inadvertently triggers a summarization capability rather than the required adjudicative reasoning.

Second, SFT fails to address this reasoning gap due to the inherent characteristics of civil judicial proceedings. Unlike general domain tasks, civil cases are characterized by high complexity and extensive length, with substantial divergence in evidence chains, factual patterns, and applicable legal statutes across different cases. This heterogeneity means that the specific logic required to adjudicate one case may differ vastly from another. Consequently, it is difficult for the model to learn generalizable features through simple fine-tuning. SFT tends to overfit the model to the specific distribution or linguistic style of the training data, but struggles to abstract the robust, transferable judicial logic required to navigate the diverse and distinct evidence patterns found in real-world civil trials.

Furthermore, we posit that the "Clerk Trap" may share a common root with known LLM behaviors such as sycophancy, largely driven by Reinforcement Learning from Human Feedback (RLHF). Models optimized for helpfulness during RLHF learn that presenting organized, balanced information is a safe, high-reward strategy, whereas making definitive decisions risks being penalized for being wrong. In the complex legal context, this training bias leads to the safe strategy of comprehensive summarization rather than the risky strategy of rigorous, evidence-based adjudication, which manifests as the "Clerk Trap".

Algorithm 1 FocalJudge

Input: Complaint C, Defense S, Evidence E, Court Debate T, Dispute Focus D_{focus}, Fact F, Laws L
Output: Generated Facts F_{gen}, Generated Reasoning R_{gen}, Generated Judgment J_{gen}, Legal Judgment Prediction P_{ljp}

```
// Fact Generation
Let Fact_Info be an empty list to store information for each focus.
for Dispute Focus d_i in D_focus do
    Let F_i ← LLM identifies plaintiff's claims, defendant's claims and undisputed facts related to focus d_i from C, S, T.
    Append the structured information (F_i) for focus d_i to Fact_Info.
end for
Let F_gen ← LLM integrates all information in Fact_Info and E to generate the final, complete facts.
// Reasoning, Judgment, and LJP
Let I_mid be an empty list for intermediate analysis.
for Dispute Focus d_i in D_focus do
    Let S_i ← LLM determines if F supports or opposes the claims under focus d_i.
    Let R_i ← LLM generates reasoning for S_i based on F and L.
    Append the analysis (S_i, R_i) for d_i to I_mid.
end for
Let P_ljp, R_gen, J_gen ← LLM predicts the overall reasoning, judgment and judgment class based on the complete I_mid.
return F_gen, R_gen, J_gen, P_ljp
```

5.5 RQ3: How to make a better use of the dispute focus?

To address RQ3, we propose FocalJudge, which is designed to resolve the dual challenges identified in RQ2: the "Clerk Trap", where LLMs fail to act as a "judge" to adjudicate evidence, and the limitations of SFT, which struggles to capture the complex, heterogeneous logic of civil cases. The framework, as shown in Algorithm 1, overcomes these problems by structurally **forcing** LLMs to decompose complex cases into key dispute focuses, aggregating relevant evidence, claims and statutes for each

Model	Fact			Reasoning			Judgment		
	ROU.	BS.	LLM	ROU.	BS.	LLM	ROU.	BS.	LLM
Small LLMs (with Basic Prompt and 1 Example)									
Gemma3	49.59 $\uparrow_{15.4}$	75.91 $\uparrow_{2.58}$	4.41 $\uparrow_{0.48}$	24.49 $\uparrow_{7.17}$	71.80 $\uparrow_{1.15}$	3.88 $\uparrow_{0.38}$	26.60 $\uparrow_{12.9}$	74.61 $\uparrow_{4.69}$	3.12 $\uparrow_{0.16}$
Llama-3.1-8B	47.54 $\uparrow_{17.2}$	75.78 $\uparrow_{1.85}$	4.25 $\uparrow_{0.38}$	26.41 $\uparrow_{8.16}$	72.21 $\uparrow_{1.37}$	3.74 $\uparrow_{0.31}$	24.59 $\uparrow_{5.47}$	70.92 $\uparrow_{5.86}$	2.78 $\uparrow_{0.20}$
Qwen3-8B	47.90 $\uparrow_{14.2}$	76.37 $\uparrow_{2.78}$	4.56 $\uparrow_{0.40}$	27.39 $\uparrow_{3.55}$	72.84 $\uparrow_{1.24}$	4.97 $\uparrow_{0.90}$	30.80 $\uparrow_{9.91}$	78.41 $\uparrow_{4.06}$	3.79 $\uparrow_{0.14}$
ChatLaw	16.10 $\uparrow_{1.85}$	67.10 $\uparrow_{3.87}$	2.65 $\uparrow_{0.29}$	17.59 $\uparrow_{4.82}$	64.51 $\uparrow_{1.68}$	2.85 $\uparrow_{0.41}$	16.93 $\uparrow_{4.80}$	57.56 $\uparrow_{2.27}$	2.76 $\uparrow_{0.06}$
LexiLaw	45.96 $\uparrow_{15.5}$	73.10 $\uparrow_{2.68}$	3.13 $\uparrow_{0.11}$	9.09 $\uparrow_{0.30}$	53.82 $\uparrow_{0.08}$	2.20 $\uparrow_{0.00}$	11.18 $\uparrow_{0.97}$	54.77 $\uparrow_{0.35}$	2.39 $\uparrow_{0.02}$
LLM APIs (with Basic Prompt and 1 Example)									
Claude-3.5	**53.48** $\uparrow_{18.5}$	**78.94** $\uparrow_{4.28}$	4.66 $\uparrow_{0.59}$	26.77 $\uparrow_{2.96}$	73.65 $\uparrow_{2.26}$	5.02 $\uparrow_{1.40}$	34.19 $\uparrow_{10.8}$	76.37 $\uparrow_{0.35}$	3.89 $\uparrow_{0.50}$
DeepSeek-v3	41.01 $\uparrow_{5.98}$	73.86 $\downarrow_{0.80}$	4.54 $\uparrow_{0.26}$	26.34 $\uparrow_{1.54}$	71.17 $\uparrow_{1.25}$	4.87 $\uparrow_{0.95}$	27.31 $\uparrow_{6.86}$	78.48 $\uparrow_{3.19}$	3.58 $\uparrow_{0.23}$
Gemini-3	51.02 $\uparrow_{17.0}$	76.51 $\uparrow_{3.13}$	**5.09** $\uparrow_{0.72}$	**32.99** $\uparrow_{5.36}$	**74.09** $\uparrow_{2.37}$	**5.56** $\uparrow_{0.49}$	34.21 $\uparrow_{10.3}$	**80.23** $\uparrow_{2.56}$	**4.07** $\uparrow_{0.48}$
Grok-4	40.70 $\uparrow_{7.22}$	75.39 $\uparrow_{1.12}$	4.61 $\uparrow_{0.40}$	28.23 $\uparrow_{3.33}$	73.18 $\uparrow_{1.86}$	5.12 $\uparrow_{0.69}$	31.47 $\uparrow_{10.3}$	79.22 $\uparrow_{3.92}$	3.85 $\uparrow_{0.77}$
GPT-4o	53.19 $\uparrow_{5.34}$	77.15 $\uparrow_{0.60}$	4.75 $\uparrow_{0.09}$	23.41 $\uparrow_{0.23}$	73.80 $\uparrow_{2.29}$	3.89 $\uparrow_{0.51}$	**38.94** $\uparrow_{12.6}$	77.00 $\uparrow_{0.69}$	3.84 $\uparrow_{0.74}$

Table 5: The Fact, Reasoning and Judgment generation results of RQ3. The best results are highlighted in bold.

Model	LJP		
	Macro F1	Weighted F1	Acc.
Small LLMs (with Basic Prompt and 1 Example)			
Gemma3	54.59 $\uparrow_{3.73}$	72.25 $\uparrow_{0.59}$	70.01 $\uparrow_{1.39}$
Llama-3.1-8B	52.48 $\uparrow_{1.74}$	72.08 $\uparrow_{1.65}$	68.60 $\uparrow_{1.96}$
Qwen3-8B	56.38 $\uparrow_{1.75}$	77.43 $\uparrow_{2.40}$	75.13 $\uparrow_{2.2}$
ChatLaw	33.97 $\uparrow_{1.55}$	63.98 $\uparrow_{1.89}$	61.87 $\uparrow_{2.27}$
LexiLaw	32.11 $\uparrow_{0.26}$	61.31 $\uparrow_{0.94}$	58.89 $\uparrow_{0.30}$
LLM APIs (with Basic Prompt and 1 Example)			
Claude-3.5	63.82 $\uparrow_{11.3}$	80.94 $\uparrow_{2.10}$	79.37 $\downarrow_{1.24}$
DeepSeek-v3	62.75 $\uparrow_{5.43}$	80.64 $\uparrow_{3.17}$	78.57 $\uparrow_{4.20}$
Gemini-3	**67.90** $\uparrow_{7.04}$	**84.73** $\uparrow_{4.54}$	**83.33** $\uparrow_{4.44}$
Grok-4	55.49 $\uparrow_{10.4}$	77.09 $\uparrow_{5.33}$	76.04 $\uparrow_{3.17}$
GPT-4o	64.93 $\uparrow_{12.1}$	81.74 $\uparrow_{2.77}$	80.61 $\downarrow_{1.30}$

Table 6: The LJP results of RQ3. The best results are highlighted in bold.

focus, and finally synthesizing a comprehensive judgment.

Fact-Finding Stage FocalJudge first directs the LLM to perform a point-by-point analysis, where for each individual dispute focus, it isolates and organizes the divergent claims of all parties alongside any undisputed facts. Only after this granular analysis is complete does it proceed to the second stage: comprehensive synthesis, where it weights conflicting claims using evidence and generates the final unified Fact determination.

Judgment Stage Subsequently, taking the dispute focus as the core, the judge makes the final judgment based on facts and laws. For generating Reasoning, Judgment, and LJP, the LLM first conducts a point-by-point adjudication, determining whether the established facts support or oppose each dispute focus and identifying the relevant laws. In the next step, it performs a final synthesis, integrating these intermediate analyses to generate the final reasoning and judgment.

Tables 5 and 6 demonstrate significant improvements over the baseline. While Claude-3.5 excels in specific Fact generation metrics, Gemini-3-Pro dominates Reasoning, Judgment, and LJP tasks, outperforming peers across nearly all evaluation criteria. We also conducted ablation experiments to demonstrate the effectiveness of our framework, which can be found in Appendix J.

5.6 Answer3: FocalJudge guides the LLM to follow a structured process, improving performance

The modest gain in Reasoning reflects the inherent complexity of judicial argumentation, which demands logic, value balancing, and statutory interpretation beyond simple information integration. While FocalJudge provides a structural architecture, professional-standard reasoning remains bounded by the model's intrinsic legal knowledge. Nevertheless, improvements across LJP, Fact and Judgment generation confirm that FocalJudge effectively leverages dispute focus to guide LLMs through complex legal scenarios.

It is worth clarifying that FocalJudge fundamentally diverges from CoT by enforcing structured judicial mapping rather than relying on open-ended reasoning. While CoT's exploratory nature risks exacerbating the "Clerk Trap", FocalJudge utilizes cognitive "dimensionality reduction" to systematically organize chaotic case data. This architectural constraint addresses the "Clerk Trap" by strictly separating claim extraction from adjudication, which simultaneously guarantees inherent interpretability. By externalizing intermediate judicial stages—such as dispute focuses and undisputed facts—FocalJudge embeds transparency di-

rectly into the workflow, rendering the decision path verifiable and distinct from post-hoc explainability methods.

6 Conclusions

In this paper, we construct FocalLaw, the first dataset to align full-process Chinese civil judicial data through the dispute focus, and systematically evaluate the role of the dispute focus in judgment prediction and the judicial proceedings. Through experiments, we discover the "Clerk Trap" and SFT limitations in civil cases. To address this gap, we propose FocalJudge, which can effectively utilize the dispute focus for legal judgment prediction, providing new insights and solutions to apply artificial intelligence technology to assist in judicial adjudication in more realistic and complex scenarios.

Limitations

Although this work makes a contribution by introducing the FocalLaw dataset and proposing the FocalJudge framework, we acknowledge several existing limitations.

First, while our proposed FocalJudge framework improves performance and provides a legally grounded reasoning structure, the improvement in generating judicial Reasoning in RQ3 is limited, indicating that the framework, while beneficial, is still constrained. High-quality judicial reasoning demands not only structure but also substantive depth, including judicial logic, value balancing, and deep legal knowledge, which may be beyond the reach of prompting alone.

Additionally, potential sample selection bias exists in the FocalLaw dataset. Our rigorous filtering process (<0.14% retention) prioritized data quality but likely excluded imperfect real-world cases, such as those with unclear dispute focuses or disorganized debates. Therefore, our conclusions are most applicable to well-structured judicial proceedings, and generalizability to more chaotic legal disputes warrants further investigation.

Furthermore, in this study, we position the dispute focus as a given input to systematically evaluate its utility in guiding judicial reasoning, rather than evaluating the model's ability to autonomously predict or generate it. In real-world judicial scenarios, distilling the dispute focus from chaotic pre-trial information is a complex cognitive task in itself. In this study, we have already successfully utilized the dispute focus; the next step is to generate the dispute focus.

Finally, regarding the monolingual nature of FocalLaw, we prioritized refining the 'dispute focus' mechanism within a single legal system. Given the lack of a unified international definition for such concepts and the significant variance in legal proceedings across regions, we consider this study a necessary prerequisite before expanding to the complexities of multilingual legal AI."

In the future, we aim to model the judicial process at a more granular level, breaking it down into finer steps to be optimized individually. Finally, there is an opportunity to expand the dataset, utilize its multi-modal content, and conduct more in-depth research on the generation and impact of the dispute focus itself.

Acknowledgements

This research is supported by the National Key R&D Program of China (grant No. 2024YFA1012700).

References

AI@Meta. 2024. Llama 3.1 model card.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Farid Ariai and Gianluca Demartini. 2025. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *Preprint*, arXiv:2410.21306.

Marco Billi, Alessandro Parenti, Giuseppe Pisano, and Marco Sanchi. 2023. Large language models and explainable law: a hybrid methodology. *Preprint*, arXiv:2311.11811.

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

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2024. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *Preprint*, arXiv:2306.16092.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Google DeepMind. 2025. Gemini 3 pro.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009, Singapore. Association for Computational Linguistics.

Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu, Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, and Fei Wu. 2019. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1361–1370, New York, NY, USA. Association for Computing Machinery.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Masaki Fujita, Takaaki Onaga, and Yoshinobu Kano. 2024. Llm tuning and interpretable cot: Kis team in coliee 2024. In *New Frontiers in Artificial Intelligence*, pages 140–155, Singapore. Springer Nature Singapore.

Gemma. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Haitao Li, Qingyao Ai, Qian Dong, and Yiqun Liu. 2024. Lexilaw: A scalable legal language model for comprehensive legal understanding.

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

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal judgment prediction with multi-stage case representation learning in the real court setting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 993–1002, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

Luís Moniz Pereira, Francisco C. Santos, and António Barata Lopes. 2024. *AI Modelling of Counterfactual Thinking for Judicial Reasoning and Governance of Law*, pages 263–279. Springer International Publishing, Cham.

Qwen. 2024. Qwen2.5: A party of foundation models.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Peizhang Shao, Linrui Xu, Jinxi Wang, Wei Zhou, and Xingyu Wu. 2025. When large language models meet law: Dual-lens taxonomy, technical advances, and ethical governance. *Preprint*, arXiv:2507.07748.

Weihang Su, Baoqing Yue, Qingyao Ai, Yiran Hu, Jiaqi Li, Changyue Wang, Kaiyuan Zhang, Yueyue Wu, and Yiqun Liu. 2025. Judge: Benchmarking judgment document generation for chinese legal system. *Preprint*, arXiv:2503.14258.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction. *Preprint*, arXiv:2212.02199.

Josef Valvoda and Ryan Cotterell. 2024. Towards explainability in legal outcome prediction models. *Preprint*, arXiv:2403.16852.

Yiwen Wang, Xiaobing Zhao, Xiaoke Qi, Bo Chen, Chuanlian Ma, and Yang Xu. 2025. A large language model evaluation method for legal case retrieval. *DATA INTELLIGENCE*, 7(2):440–460.

Bin Wei, Yaoyao Yu, Leilei Gan, and Fei Wu. 2025. An llms-based neuro-symbolic legal judgment prediction framework for civil cases. *Artificial Intelligence and Law*.

xAI. 2025. Grok 4. https://x.ai/news/grok-4.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *Preprint*, arXiv:1807.02478.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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

A Dataset Sample

Figure 5 shows a random sample selected from the FocalLaw dataset. Each extracted field corresponds to the explanation in Figure 2.

B Prompt for Structuring Legal Case Documents

The initial and most critical phase in the construction of the FocalLaw dataset involves transforming raw, semi-structured legal case documents into a highly structured, machine-readable format. To achieve this, a detailed and comprehensive prompt is designed to guide an LLM through the extraction process, as shown in Figure 6. This prompt serves as the foundation for the initial automated annotation step, which is later followed by a rigorous manual review to ensure data accuracy.

The prompt instructs the LLM to act as an extractor, processing judicial documents and outputting key information in a structured JSON format. The core principles guiding this task are original text extraction, completeness, and accuracy. This meant the LLM is explicitly directed to pull content directly from the source text without any modification, summarization, or rephrasing, ensuring the integrity and fidelity of the original legal language.

The prompt provides specific instructions for extracting 10 distinct fields that represent the entire judicial process, including Case Number, Complaints(plaintiff's claims), Defense, Facts as Determined, Court Reasoning, Judgment, Cited Laws, and Evidence. Crucially, the prompt includes highly specific guidance for identifying and extracting the Dispute Focus, which is central to the research. It directs the model to search for explicit markers like "Dispute Focus" or "Focus" within the "This court holds" section of the document and to preserve the concise, typically numbered format of these key points of contention. This level of detail is essential for systematically isolating the core legal questions that structure the judicial proceedings, thereby creating a reliable foundation for the subsequent experimental analysis.

C Prompts for Comparative Experiments

The prompts used in our experiments are developed through an iterative refinement process. We design the prompts to be as simple as possible because the prompting technique is not the main point of our research. Key design principles include: 1) Role-Playing, where the LLM is instructed to act as a 'professional legal assistant' to adopt the appropriate analytical perspective; 2) Instructions, particularly within the FocalJudge framework, to guide the model and to do the required tasks. This design ensures that prompts are clear, unambiguous, and do not contain extra task prompts; and 3) Structured Output, where prompts specify the exact format for the response to ensure consistency.

In this study, we focus exclusively on prompting-based methods to assess the inherent, out-of-the-box reasoning capabilities of existing LLMs without model-specific modifications. This approach allows us to benchmark a wide range of models on their ability to understand and apply complex legal frameworks and the dispute focus in a one-shot setting. The prompt used in RQ1 to 3 is shown in Figure 2 to 6.

Case Study: (201) Chuan 1602 Min Chu No.

Pre-Trial Information

Cause: Sales Contract Dispute

Plaintiff's Claims: 1. Revoke the "Domestic Car Sales Contract". 2. Refund the car purchase price of 250,000 RMB. 3. Compensation for expenses (insurance, tax, etc.) totaling 42,442.68 RMB. 4. Treble damages (750,000 RMB) for alleged fraud. 5. Compensation for loan interest losses (7,391.83 RMB). *Grounds:* The plaintiff discovered severe damage to the undercarriage and mudguards after purchase. The plaintiff alleges the defendant concealed this fact, constituting fraud.

Defendant's Defense: 1. No grounds exist to revoke the contract. 2. The vehicle delivered was inspected and confirmed by the plaintiff to be free of defects. 3. The defendant had no intent to defraud and committed no fraudulent acts; treble damages should be rejected.

Key Evidence Submitted:

- **"Domestic Car Sales Contract":** Signed by the Plaintiff and Defendant on July 18, 201, establishing the purchase agreement.
- **"Vehicle Delivery Inspection Confirmation":** Signed by the Plaintiff on Sept 15, 201, stating "exterior and interior are in good condition, no scratches or damage".
- **"Judicial Appraisal Report":** Issued by Forensic Center, stating the undercarriage damage was caused by scraping against a hard object but did not affect safety performance.

Dispute Focus

1. Did the defendant's conduct constitute fraud? Should the defendant pay treble damages, and should the contract be revoked?
2. Should the plaintiff's request for a refund of 250,000 RMB and compensation for related losses be supported?

Court Debate

Plaintiff: The surveillance video from AM on delivery day clearly shows a **suspicious object hanging** from the vehicle's undercarriage. The damage existed *before* handover. The defendant claims to have inspected the car twice (July and Aug). If they put the car on a lift, the damage would have been obvious. They must have known.

Defendant: First, the "hanging object" in the video is unclear; it could be a shadow or light reflection. Second, regarding the inspection: According to the *Passenger Car Pre-delivery Inspection (PDI) Guidelines* by the China Automobile Dealers Association, the standard for exterior inspection is **visual observation within 1 meter**. It does not require crawling under the car or lifting it. We followed the industry standard, and no damage was visible at that time.

Plaintiff: Does "within 1 meter" only apply to the paint? If the paint is perfect but the chassis is shattered, is it still a "perfect new car"? The defendant's argument is invalid.

After Trial Information

Facts: The Plaintiff picked up the car on Sept 15, 201, signing a confirmation that the "exterior and interior are in good condition." On Sept 21, damage to the undercarriage was discovered at the Defendant's shop. Judicial appraisal confirmed the damage was caused by scraping against a hard object but did not affect safety. The specific time of damage (pre- or post-handover) could not be determined.

Reasoning:

- **On Fraud (Focus 1):** The Plaintiff signed the delivery confirmation. There is insufficient evidence to prove the damage existed *before* handover or that the Defendant intentionally concealed it. Thus, fraud is not established, and the claim for treble damages is rejected.
- **On Refund (Focus 2):** Under the *Law on Protection of Consumer Rights and Interests*, for defects found within 6 months, the burden of proof lies with the seller. The Defendant failed to prove the damage occurred *after* delivery. Since the Plaintiff cannot achieve the contract purpose of purchasing a "new car," the contract can be terminated.

Laws: *Consumer Rights Protection Law* Art. 23(3); *Contract Law* Art. 94(4), 97.

Judgment: 1. Defendant shall refund 250,000 RMB; Plaintiff shall return the vehicle. 2. Defendant shall pay 42,442.68 RMB for expenses (tax, insurance, etc.). 3. Defendant shall pay 7,391.83 RMB for interest losses. 4. Other claims (treble damages) are dismissed.

Judgment Class: 2 (Partial Support)

Figure 5: An example case from the FocalLaw dataset (Case Code: (201) Chuan 1602 Min Chu No.). The content is translated into English and structured into Pre-Trial, Dispute Focus, and After Trial phases.

Task Description
Please structurally process the input judicial documents, extract key information and output it in JSON format.
Extraction Requirements
Basic Principles
1. **Original Text Extraction:** All extracted content must be directly taken from the original text without any processing, modification, or summarization.
2. **Completeness:** Ensure that the extracted content is complete and does not omit any important information.
3. **Accuracy:** Extract strictly according to the original text content and maintain the original expression style.
Field Extraction Instructions
1. Case Number (code): Extract the case number, such as "(2018) Xiang 0104 Civil Initial 8979"
2. Complaint: Extract the complete content of the plaintiff's claims and factual grounds.
3. Defense: Extract the defendant's defense content, including the defense opinions of all defendants.
4. Establishing the Facts: Extract the content of the factual determination following the phrase "Upon examination and verification" or similar expressions.
5. Dispute Focus (dispute_focus): Stored in dictionary format
- key: "0", "1", "2" and other numeric strings
- value: Specific content of the dispute focus (in short sentence form)
- Extraction Location: Mainly search in the "This court holds" section and its vicinity, usually with obvious identifiers such as "Dispute Focus" or "Focus"
- Format Characteristics:
- Dispute focuses are typically short sentences, such as "Regarding the nature of the contract", "Regarding the calculation of the principal and interest owed", etc.
- Common formats: (1), (2), (3) or 1, 2, 3 and other numbered identifiers
- Example:
- "Regarding the nature of the contract"
- "Regarding the calculation of the principal and interest owed"
- "Regarding the determination of the responsibility of the defendant, Hunan Houxin E-commerce Co., Ltd."
- Extraction Method:
- Prioritize searching for content explicitly marked as "Dispute Focus"
- Next, search for numbered short sentence titles in the "This court holds" section
- Ensure to extract the complete expression of the dispute focus but maintain a concise short sentence form
- Do not extract uncertain dispute focuses
6. Court Reasoning: Extract the complete content of the part starting with “The court holds" (This Court holds that
7. Judgement Result: Extract the main body of the judgement (usually following "The judgement is as follows")
8. Judgment Result Classification (judgment_class): Array format
- 0: All of the plaintiff's requests are rejected
- 1: All of the plaintiff's requests are accepted
- 2: Some of the plaintiff's requests are accepted
- Classification is based on the judgment result and can be a single value or an array of multiple values.
9. Cited Laws: Stored in a dictionary format
- Key: Name of the law, such as "Article 60 of the Contract Law of the People's Republic of China"
- Value: Specific content of the law
- Include all laws cited in the judgment.
10. Evidence in documents: Stored in dictionary form
- Key: Name of the evidence, such as "Investment Contract", "WeChat chat records", etc.
- Value: Specific description or content of the evidence
Output Format
```json
{
"case_number": "Case Number Content",
“complaint": “Complaint Content",
"defense": "Defense Content",
"fact": "Facts as Determined Content" "dispute_focus": {
"0": "The first point of contention",
"1": "The second point of contention" },
"reasoning": "Court reasoning content",
"judgement": "Judgement result content" "judgment_class": [0/1/2],
"laws": {
"Name of the law and its articles": "Specific content of the articles" },
"evidences": {
"Evidence Name": "Evidence Content and Description" }
}
```

Figure 6: Prompt for the first step in structuring legal case documents.

You are a professional legal assistant, required to generate judgment reasoning and predict the verdict based on the provided case information.
/* Example */
(Reference case)

/* Cases to be Predicted */
Cause of action:
Plaintiff's claim:
Defendant's defense:
Court Debate:
Evidence:
Dispute Focus: **(If the dispute focus is included)**

/* Instructions */
Please complete the following tasks based on the above reference cases and the information of the case to be predicted:
1. Predict the verdict(Please provide detailed reasoning for the judgment)
2. Predict the type of verdict (choose one):
- 0: The plaintiff's claim was rejected
- 1: The plaintiff's request is accepted
- 2: The plaintiff's claim was partially accepted
3. Output the reasoning behind the judgment (the reasoning process of the judgment)

Please output in the following format:
<predicted_judgement>
(Detailed content of the judgment)
</predicted_judgement>

<predicted_class>
(Number 0, 1, or 2)
</predicted_class>

<judgement_reasoning>
(Please provide a detailed description of the court's judgment based on the focus of the dispute)
(Please provide detailed reasoning for the judgment)
</judgement_reasoning>

Figure 7: Prompt used in RQ1.

You are a professional legal assistant who needs to generate judgment reasoning and predict the verdict based on the provided case information.
/* Example */
(Reference case)

/* Cases to be Predicted */
Cause of action:
Plaintiff's claim:
Defendant's defense:
Court Debate:
Evidence:
Dispute Focus: **(If the dispute focus is included)**

/* Instructions */
Please generate the facts ascertained by the court based on the above reference cases and the information of the case to be predicted.
requirement:
1. The facts generated should be objective, accurate, and detailed
2. The plaintiff's claims, the defendant's defense, the court trial summary, detailed evidence, and information from the event content should be comprehensively considered

Please output in the following format:
<predicted_fact>
(Detailed court findings of fact)
</predicted_fact>

Figure 8: Prompt used in the Fact generation task in RQ2.

You are a professional legal assistant, required to generate judgment reasoning and predict the verdict based on the provided case information.
/* Example */
(Reference case)

/* Cases to be Predicted */
Cause of action:
Plaintiff's claim:
Defendant's defense:
Fact:
Laws:
Dispute Focus: **(If the dispute focus is included)**

/* Instructions */
Please complete the following tasks based on the above reference cases and the information of the case to be predicted:
1. Predict the verdict (Please provide detailed reasoning for the judgment)
2. Predict the type of verdict (choose one):
- 0: The plaintiff's claim was rejected
- 1: The plaintiff's request is accepted
- 2: The plaintiff's claim was partially accepted
3. Output the reasoning behind the judgment (the reasoning process of the judgment)

Please output in the following format:
<predicted_judgement>
(Detailed content of the judgment)
</predicted_judgement>

<predicted_class>
(Number 0, 1, or 2)
</predicted_class>

<judgement_reasoning>
(Please provide a detailed description of the court's judgment based on the focus of the dispute)
(Please provide detailed reasoning for the judgment)
</judgement_reasoning>

Figure 9: Prompt used in Reasoning, Judgment, and LJP tasks in RQ2.

// First call to LLM //
You are a professional legal assistant, required to generate judgment reasoning and predict the verdict based on the provided case information.
/* Example */
(Reference case)

/* Cases to be Predicted */
Cause of action:
Plaintiff's claim:
Defendant's defense:
Court Debate:
Evidence:
Dispute Focus: **(If the dispute focus is included)**

/* Instructions */
1.The generated facts should be objective, accurate, and detailed.
2.The information from the plaintiff's claims, the defendant's defense, the court session summary, detailed evidence, and the event description should be comprehensively considered.
3.For undisputed facts, please state them directly as declarative sentences. For disputed facts, please objectively describe the core claims and actions of both parties.
Explanation of Extracted Fields:

For each point of dispute
plaintiff_claims: The plaintiff's core claims and actions regarding this point of dispute.
defendant_claims: The defendant's core claims and actions regarding this point of dispute.
undisputed_facts: Facts that are not disputed by either party, stated as declarative sentences.
Example of Extraction Format:
```
{
  "Name of Point of Dispute 1": {
    "plaintiff_claims": ["Transferred 100,000 RMB via XX Bank on 2023-05-01 (Voucher No.: TT2023050112)"
    "defendant_claims": ["Acknowledges receipt of payment, but claims it was for repayment of goods"
    "undisputed_facts": ["The defendant's account received 100,000 RMB"]
  },
  "Name of Point of Dispute 2": {...}
}
```
// Second call to LLM //
Evidence:
Now, please refer to the materials and specific evidence to generate the ascertained facts in a style appropriate for a legal instrument.

Figure 10: Prompt used in the Fact generation task in RQ3.

// First call to LLM //
You are a professional legal assistant, required to generate judgment reasoning and predict the verdict based on the provided case information.
/* Example */
(Reference case)

/* Cases to be Predicted */
Cause of action:
Plaintiff's claim:
Defendant's defense:
Fact:
Laws:
Dispute Focus: **(If the dispute focus is included)**

/* Instructions */
The description of the extracted fields is as follows: For each point of dispute:
1.fact (A fact from the court-ascertained facts, which needs to be clear and complete)
2.relation (The relationship with the point of dispute, fill in 'support' or 'oppose')
3.reasoning (The analytical reasoning based on the "court-ascertained facts").
4.laws (The legal provisions most relevant to the fact)
The extraction format example is as follows:
{
"Name of Point of Dispute 1": [
{"fact": "Bank transfer of 100,000 yuan on 2023-05-01",
"relation": "support",
"reasoning": "The court-ascertained facts confirm the authenticity of the transfer, which constitutes preliminary evidence of the payment."
"laws":},
{"fact": "Procurement Contract PC-0415",
"relation": "oppose",
“reasoning”: “The court-ascertained facts indicate that the contract does not directly correspond to the transfer in terms of time and amount.
"laws":...}
],
"Name of Point of Dispute 2": [...],
}
// Second call to LLM //
Please complete the following tasks based on the above reference cases and the information of the case to be predicted:
1. Predict the verdict (Please provide detailed reasoning for the judgment)
2. Predict the type of verdict (choose one):
- 0: The plaintiff's claim was rejected
- 1: The plaintiff's request is accepted
- 2: The plaintiff's claim was partially accepted
3. Output the reasoning behind the judgment (the reasoning process of the judgment)

Please output in the following format:
<predicted_judgement>
(Detailed content of the judgment)
</predicted_judgement>

<predicted_class>
(Number 0, 1, or 2)
</predicted_class>

<judgement_reasoning>
(Please provide detailed reasoning for the judgment)
</judgement_reasoning>

Figure 11: Prompt used in Reasoning, Judgment, and LJP tasks in RQ3.

D Detailed Information of the Fine-tuning

We conduct fine-tuning using our custom training pipeline. To avoid overfitting, we monitor the trends of training and test losses and confirm through preliminary experiments that the performance stabilizes after 3 epochs. Therefore, all models are trained for 3 epochs. We adopt the LoRA method and evaluate the results through ten-fold cross-validation. The hyperparameters are presented in Table 7.

Hyperparameters	Value
Epochs	3
Batch size	1
Learning rate	5e-5
Compute type	fp16
Gradient accumulation	8
Maximum gradient norm	1.0

Table 7: Detailed hyperparameter settings.

E Details and Validation of the LLM Evaluation Metric

During the evaluation, we report the BERTscore. Since FocalLaw is a Chinese legal dataset, BERTscore is initialized using chinese-bert-wwm [‡].

To ensure the reliability of LLM evaluation, we conduct a human correlation study to validate whether the scores assigned by the LLM judge (GPT-4o) align with those from human legal experts. We randomly select a subset of 100 generated output from the Reasoning generation task. These outputs are sourced proportionally from the different models evaluated in our experiments. Our annotation team consists of three experts with legal background who underwent systematic training. The rating criteria are the same as the LLM-as-a-judge prompt proposed in (Li et al., 2025). The results, summarized in Table 8, show a strong correlation for both metrics. While not a perfect substitute for expert legal review, this result supports its use as a scalable and consistent evaluation metric throughout our experiments.

Metrics	LLM Score
Pearson	0.625
Spearman	0.714

Table 8: Correlation between LLM-as-Judge and Human Expert Scores

[‡] https://huggingface.co/hfl/chinese-bert-wwm

Model	Fact			Reasoning			Judgment		
	ROU.	BS.	LLM	ROU.	BS.	LLM	ROU.	BS.	LLM
Small LLMs (with Basic Prompt and 1 Example)									
Gemma3	33.48 $\downarrow_{0.72}$	73.20 $\downarrow_{0.13}$	3.43 $\downarrow_{0.50}$	18.04 $\uparrow_{0.72}$	71.56 $\uparrow_{0.91}$	3.69 $\uparrow_{0.19}$	13.35 $\downarrow_{0.33}$	68.45 $\downarrow_{1.47}$	2.76 $\downarrow_{0.20}$
Llama-3.1-8B	29.87 $\downarrow_{0.44}$	71.20 $\downarrow_{1.28}$	3.48 $\downarrow_{0.39}$	18.81 $\uparrow_{0.56}$	72.25 $\uparrow_{1.41}$	3.52 $\uparrow_{0.09}$	17.80 $\downarrow_{1.32}$	63.42 $\downarrow_{1.64}$	2.25 $\downarrow_{0.33}$
Qwen3-8B	31.97 $\downarrow_{1.74}$	72.83 $\downarrow_{0.76}$	3.94 $\downarrow_{0.22}$	24.76 $\uparrow_{0.92}$	72.81 $\uparrow_{1.21}$	4.24 $\uparrow_{0.17}$	20.61 $\downarrow_{0.28}$	71.50 $\downarrow_{2.85}$	3.42 $\downarrow_{0.23}$
ChatLaw	14.21 $\downarrow_{1.26}$	62.88 $\downarrow_{0.35}$	2.30 $\downarrow_{0.06}$	12.83 $\uparrow_{0.06}$	63.73 $\uparrow_{0.90}$	2.49 $\uparrow_{0.05}$	12.30 $\uparrow_{0.17}$	55.27 $\downarrow_{0.02}$	2.70 $\downarrow_{0.04}$
LexiLaw	29.32 $\downarrow_{1.11}$	70.08 $\downarrow_{0.34}$	3.19 $\downarrow_{0.05}$	7.94 $\downarrow_{0.85}$	53.30 $\downarrow_{0.44}$	2.17 $\downarrow_{0.03}$	10.47 $\uparrow_{0.26}$	53.81 $\downarrow_{0.61}$	2.32 $\downarrow_{0.05}$
LLM APIs (with Basic Prompt and 1 Example)									
Claude-3.5	34.30 $\downarrow_{0.73}$	74.25 $\downarrow_{0.41}$	3.94 $\downarrow_{0.13}$	25.83 $\uparrow_{2.02}$	72.94 $\uparrow_{1.55}$	3.90 $\uparrow_{0.10}$	23.41 $\uparrow_{0.03}$	74.23 $\downarrow_{1.79}$	3.32 $\downarrow_{0.07}$
DeepSeek-v3	33.02 $\downarrow_{1.01}$	72.86 $\downarrow_{0.52}$	4.12 $\downarrow_{0.16}$	25.00 $\uparrow_{0.20}$	71.20 $\uparrow_{1.28}$	3.95 $\uparrow_{0.03}$	21.09 $\uparrow_{0.64}$	72.82 $\downarrow_{2.47}$	3.03 $\downarrow_{0.32}$
Gemini-3	37.36 $\downarrow_{0.57}$	74.25 $\downarrow_{0.21}$	4.26 $\downarrow_{0.11}$	**29.65** $\uparrow_{2.02}$	**73.95** $\uparrow_{2.23}$	**5.31** $\uparrow_{0.24}$	24.84 $\uparrow_{0.97}$	77.75 $\uparrow_{0.08}$	**3.74** $\uparrow_{0.15}$
Grok-4	32.63 $\downarrow_{0.85}$	73.92 $\downarrow_{0.35}$	4.17 $\downarrow_{0.04}$	28.07 $\uparrow_{3.17}$	73.15 $\uparrow_{1.83}$	4.58 $\uparrow_{0.15}$	22.34 $\uparrow_{1.17}$	74.35 $\downarrow_{0.95}$	3.25 $\uparrow_{0.17}$
GPT-4o	**46.93** $\downarrow_{0.92}$	**76.08** $\downarrow_{0.47}$	**4.48** $\downarrow_{0.18}$	25.70 $\uparrow_{2.52}$	73.31 $\uparrow_{1.80}$	3.64 $\uparrow_{0.26}$	**25.05** $\downarrow_{1.31}$	**79.00** $\uparrow_{0.04}$	3.05 $\downarrow_{0.05}$

Table 9: The Fact, Reasoning and Judgment generation results of RQ2. The best results are highlighted in bold.

Model	LJP		
	Macro F1	Weighted F1	Acc.
Small LLMs (with Basic Prompt and 1 Example)			
Gemma3	49.61 $\downarrow_{1.25}$	70.75 $\downarrow_{0.91}$	69.28 $\uparrow_{0.66}$
Llama-3.1-8B	48.10 $\downarrow_{2.64}$	69.14 $\uparrow_{1.29}$	68.20 $\uparrow_{1.56}$
Qwen3-8B	49.66 $\downarrow_{4.97}$	72.62 $\downarrow_{2.41}$	70.35 $\downarrow_{2.51}$
ChatLaw	32.72 $\uparrow_{0.30}$	62.34 $\uparrow_{0.25}$	60.40 $\uparrow_{0.80}$
LexiLaw	31.59 $\downarrow_{0.26}$	60.18 $\downarrow_{0.19}$	58.43 $\downarrow_{0.16}$
LLM APIs (with Basic Prompt and 1 Example)			
Claude-3.5	51.01 $\downarrow_{1.48}$	78.36 $\downarrow_{0.48}$	79.35 $\downarrow_{1.26}$
DeepSeek-v3	53.48 $\downarrow_{3.84}$	74.80 $\downarrow_{2.66}$	71.36 $\downarrow_{3.01}$
Gemini-3	**66.33** $\uparrow_{5.47}$	**82.60** $\uparrow_{2.41}$	**82.92** $\uparrow_{4.03}$
Grok-4	50.15 $\uparrow_{5.07}$	76.35 $\uparrow_{4.59}$	75.90 $\uparrow_{3.03}$
GPT-4o	50.45 $\downarrow_{2.42}$	77.90 $\downarrow_{1.07}$	79.90 $\downarrow_{2.01}$

Table 10: The LJP results of RQ2. The best results are highlighted in bold.

F Detail Comparative Experiment Results of RQ2

Based on the experimental results in Table 9 and 10, directly informing LLMs the dispute focus leads to a counter-intuitive and unexpected decrease in performance across several tasks. The models' performance for Fact generation shows a consistent slight drop. Additionally, the models' ability to generate the final Judgment and perform LJP declines when they are supplied with the dispute focus, even if the performance of Reasoning improves.

G Detail CoT and SFT results in RQ2

G.1 SFT Results

In this section, we provide a detailed analysis of the Supervised Fine-Tuning (SFT) results for the Gemma model, supplementing the comparative experiments discussed in Section 5.3.

To further investigate the limitations of standard optimization techniques in civil judicial scenarios, we conducted SFT experiments on Gemma3-12B using the same training pipeline and hyperparameters as applied to Qwen3-8B (see Appendix D). The performance comparison across Fact, Reasoning, Judgment generation, and LJP tasks is illustrated in Figure 12.

Consistent with the findings in the main text, SFT generally fails to yield significant performance gains for Gemma, reinforcing the "Clerk Trap" hypothesis where models struggle to transition from summarization to adjudicative reasoning even after fine-tuning.

Comparative Analysis with Qwen: When comparing the fine-tuning effects of Gemma3-12B against Qwen3-8B, we observe two distinct trends:
• Overall Performance: In terms of absolute performance metrics, Gemma's SFT results are generally inferior to those of Qwen. This aligns with the baseline capabilities of the two models observed in Table 9 and Table 10.
• Metric-Specific Resilience: Despite the lower overall performance, Gemma exhibits a slightly better comparative effect on certain metrics after fine-tuning. While Qwen experienced a consistent decline across all metrics after SFT, Gemma demonstrated a slight improvement in Reasoning and ROUGE-L in Judgment.

This nuance suggests that while neither model could fully overcome the complexity of civil adjudication through simple SFT, the impact of fine-tuning varies across model architectures, with Gemma showing different sensitivity patterns in specific sub-tasks.

G.2 CoT Results

In an attempt to address the counter-intuitive performance decline observed in RQ2, where simply

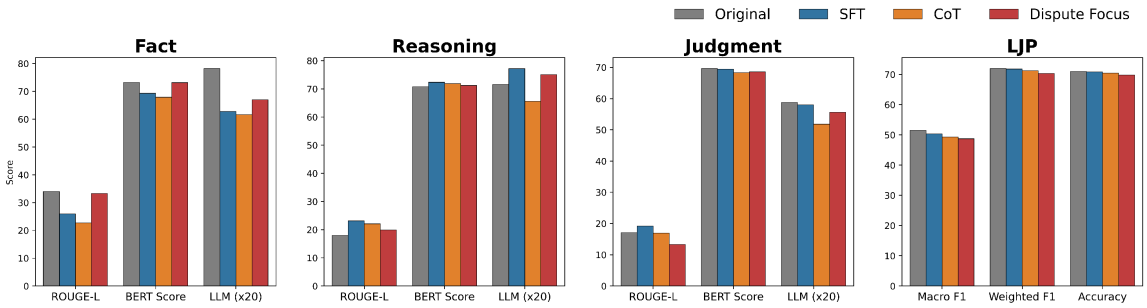

Figure 12: Detailed experimental results of Gemma3 in RQ2 across Fact, Reasoning, Judgment, and LJP tasks. The chart compares the performance of Original, SFT, and COT strategies.

/* Requirements */
Please strictly play the role of a judge and analyze and output according to the following thought process

Step 1 (Identifying Core Conflicts): Thoroughly analyze the core claims of the plaintiff and defendant and the focus of the case dispute.

Step 2 (Sorting out the nodes of the evidence chain): Based on the records of the trial evidence and cross examination process, identify the key evidence submitted by all parties (including documentary evidence/electronic data/witness testimony), with the focus of the dispute, mark the three controversial points of the evidence and their evidential power confrontation.

Step 3 (Positioning the Focus of Legal Application): Referring to the dispute focus, extract the points of legal application disagreement during the court investigation stage regarding the characterization of legal relationships, interpretation of terms, and determination of fault. Pay special attention to the focus of the judge's review of the correspondence between facts and legal elements.

Step 4 (Integrated Output): Rewrite the above analysis into a fluent and focused factual document. Please focus on the core of the dispute and cleverly connect the logic of evidence, defenses, and other content.

Figure 13: The CoT prompt used in experiments for RQ2.

providing the dispute focus proved counterproductive, alternative prompting strategies are explored. A CoT style prompt, as shown in Figure 13, is designed.

Our CoT was not designed arbitrarily, but to follow a four-step process: identify core conflicts, sort the evidence chain, position the legal application, and then synthesize this analysis into a final output. It's worth mentioning that, the CoT attempts to perform "reasoning," whereas FocalJudge performs a "structural mapping of judicial procedures," which is the cognitive "dimensionality reduction" mentioned in the paper. The cognitive goals of the two are essentially different.

However, this approach does not solve the underlying issue; instead, it results in an even more significant degradation of performance. In experiments using capable models such as GPT-4o and Qwen3-8B, the application of this CoT technique leads to a sharp and consistent drop across all evaluation metrics. Specifically, the ROUGE-L score for all text generation tasks decrease by over 5%, the BERTscore drop between 1% and 5%, and the LLM judge score decreases by 0.1 to 0.4 points. Most strikingly, the Macro F1 score for the LJP task decreases by more than 10%.

The CoT prompt, by forcing the LLM to focus even more intensely on the dispute focus, may exacerbate its inability to distinguish between conflicting "claims" and objective "facts". Instead of helping the model act as a judge, the technique seemingly reinforces its role as a "clerk", adept at summarizing disputes but incapable of resolving them to determine factual truth.

H Comparative Case Study on Fact Generation

To intuitively demonstrate the "Clerk Trap" and the impact of different strategies on the *Fact Generation* task, we present a comparative case study based on a traffic accident liability dispute.

Table 11 compares the narrative style and content handling of the Ground Truth against the outputs from the Original (Baseline), SFT, Dispute Focus, and CoT methods.

H.1 Analysis of the "Clerk Trap"

The comparison reveals a fundamental shift in cognitive mode triggered by the introduction of the dispute focus:

- **Ground Truth & Original (The Judge):** The

Method	Narrative Stance	Handling of Medical Fees (Example)	Handling of Disputes
Ground Truth	**The Judge (Decisive):** Objectively states established facts, dates, and specific amounts ascertained by the court.	*"Incurred hospitalization and examination fees totaling 112,251.53 RMB."* (Stated as a final fact).	Disputes are resolved. It states the final disability grade and insurance status as facts, without listing the arguments.
Original	**The Judge (Decisive):** Mimics the judicial tone. States facts directly without attributing them to claims.	*"Plaintiff incurred medical expenses totaling 112,361.53 RMB."* (Stated as fact).	Generally ignores the back-and-forth arguments, focusing on the result (responsibility division, insurance coverage).
SFT	**Hybrid (Inconsistent):** Starts as a judge but drifts into listing claims towards the end.	*"Plaintiff claims medical fees 112,361.53 RMB..."* (Shifts to describing the claim).	Explicitly lists the defendant's objections: *"Insurance company objects to... and applies for witness testimony."*
Dispute Focus	**The Clerk (Summarizing):** Heavily focuses on summarizing the *positions* of both parties regarding the focal points.	*"Plaintiff claims medical fees 112,361.53 RMB... Defendant argues non-medical insurance parts should be deducted."*	Devotes significant space to the defendant's specific objections to the disability grade and fee calculations, treating them as facts of the trial rather than resolving them.
CoT	**The Clerk (Process-Oriented):** Lists evidence and arguments to "prove" the case, rather than stating the outcome.	*"Plaintiff asserts... Defendant argues..."*	Lists the evidentiary documents submitted to prove the status, rather than stating the status itself.

Table 11: Comparison of Fact Generation strategies. The "Clerk Trap" is evident in the Dispute Focus and CoT methods, where the model shifts from ascertaining facts to summarizing conflicting claims.

Ground Truth establishes the timeline, diagnosis, and costs as objective realities. For example, it explicitly lists the parents' birth dates to justify the "dependents' living expenses" (though not explicitly calculated in the fact section, the basis is laid). The **Original** model, unaware of the specific dispute focus, attempts to mimic this declarative style. It successfully states: *"The accident occurred on August 19... Guo bears main responsibility."*

- **Dispute Focus (The Clerk):** When provided with the dispute focus (e.g., *"Is the loss calculation according to law?"*), the model stops acting as a decision-maker. Instead of stating *"The medical fee is 112,251.53 RMB,"* it pivots to a summary of the conflict: *"The plaintiff claims 112,361.53 RMB... The defendant argues that non-medical insurance parts should be deducted."*

  This is the essence of the **Clerk Trap**. The model interprets the presence of a Dispute Focus as an instruction to *describe the dispute* rather than *resolve the facts*. It accurately summarizes the courtroom debate (what a clerk does) but fails to filter this information into a set of legally ascertained facts (what a judge does).

- **SFT & CoT Failure: SFT** fails to correct this because the training data contains heterogeneous case structures; the model learns to mention the insurance objection but fails to integrate it into a final ruling. **CoT** exacerbates the issue. By forcing the model to "analyze evidence chains" and "identify conflicts" (Step 1 & 2 of the CoT prompt), it reinforces the focus on the *process* of the argument. The resulting output is a detailed list of who submitted what evidence and what they argued, which effectively pollutes the "Fact" section with unresolved allegations.

This case study confirms our hypothesis in RQ2: directly injecting the dispute focus into the input context without a structured adjudication framework (like FocalJudge) confuses the LLM's role, causing it to revert to a summarization task and degrading the quality of Fact Generation.

I Case Study of Failure Examples of RQ2

The first case study focuses on the Fact generation, as shown in Figure 14. The yellow background color indicates that this fact is mentioned both in the results with and without the dispute focus. The cyan background color indicates that this fact is only mentioned in the results without the dispute focus. This case shows that LLMs get counterproductive when informed of the dispute focus when generating Fact.

When tasked with generating the facts of the case without the dispute focus, the model performs normally. The output lists specific, undisputed details that form the foundation of the case, such as the exact contract names and entities (Authorization Letter for Entrustment of Leasing Commercial Properties), the full legal names of the companies involved ("Ningxia xxx Co., LTD."), and the precise monetary value of a previous court judgment (17,328,990 yuan). This output successfully mirrors the structure of an actual court document, separating the established background from the claims of the opposing parties. In contrast, when the model is provided with the dispute focus, its performance degrades. The task paradigm shifts from objective fact-finding to targeted conflict summarization. The model over-focuses on the dispute focus and omits the very undisputed background details it had previously identified, such as the specific judgment amount. Furthermore, the tone becomes argumentative rather than declarative, describing what a party "claimed" or "questioned" instead of what is factually ascertained. This demonstrates the model's inability to distinguish between an "unresolved dispute" and an "established fact," leading it to summarize the arguments instead of generating the factual basis required for a legal judgment.

The second case study focuses on Reasoning generation and LJP, as shown in Figure 15. The yellow background color indicates the dispute focus. The cyan background color indicates the critical facts or articles of law when making a judgment.

In this instance, the model, both with and without the dispute focus, incorrectly determines that a private loan relationship is established between the plaintiff and the defendant. Based on this flawed premise, it wrongly concludes that the defendant should repay the principal of 120,000 yuan with interest. While providing the dispute focus helps the model produce a more legally rigorous reasoning structure compared to the reasoning without the dispute focus, it ultimately fails to reach the correct conclusion. This failure stems from its inability to effectively utilize the most crucial facts of the case or cite the correct articles of law, a shortcoming present in both with and without the dispute focus. Specifically, the model completely misses the important third party, incorrectly assuming a simple loan relationship and failing to cite the specific legal regulation on the burden of proof that is central to the actual judgment. Furthermore, some critical information, including the acknowledgment of the partnership, emerges only during the trial proceedings, highlighting that access to court debate data is indispensable for accurate LJP.

J Ablation Study: Effectiveness of Dispute Focus in FocalJudge

To address the question of whether the performance improvement of FocalJudge stems from the multi-step prompting structure itself or the explicit injection of the *Dispute Focus*, we conducted an ablation study using GPT-4o.

J.1 Experimental Setup

We designed a variant of the framework named **FocalJudge (w/o Gold Focus)**. In this setting, we removed the ground-truth Dispute Focus from the input. Instead, we instructed the LLM to first generate the dispute focus autonomously based on the pre-trial information, and then proceed with the subsequent steps of the FocalJudge architecture. This setup isolates the contribution of the "Gold Dispute Focus" while preserving the "Multi-step Structure".

J.2 Results and Analysis

The comparative results between the standard FocalJudge (with Gold Focus) and the ablation variant are summarized in Table 12.

Metric	FocalJudge	w/o Gold Focus	Decline
ROUGE-L	**Baseline**	$\sim$ Lower	$\downarrow$ 5.0% - 9.0%
BERT-Score	**Baseline**	$\sim$ Lower	$\downarrow$ 3.0% - 5.0%
LLM Score	**Baseline**	$\sim$ Lower	$\downarrow$ 0.3 - 0.6

Table 12: Ablation study results on GPT-4o. Removing the ground-truth dispute focus causes a significant performance drop across all tasks.

Analysis and Conclusion: The experimental results demonstrate a significant degradation in performance when the model is forced to generate the dispute focus itself. This validates our hypothesis that the effectiveness of FocalJudge does not solely arise from the "chain-like" structure. Instead, the *Dispute Focus* serves as a critical cognitive anchor. The structure is merely the mechanism to leverage this anchor. Without the precise guidance of the Dispute Focus, the multi-step structure fails to guide the model effectively, leading to the "Clerk Trap" where the model summarizes rather than adjudicates. Thus, the *Dispute Focus* and the *FocalJudge framework* are inseparable.

With Dispute Focus

In this case, the core dispute lies in whether the defendant, Ningxia [redacted] Co., LTD., should pay the plaintiff, Cai, the rent loss it has incurred against Song. On May 16, 2014, Song signed a "Shop Rental Entrustment Agreement" with the defendant, entrusting the shop to the defendant for rental, but the defendant failed to pay the rent. In 2016, the court ruled that Beijing [redacted] Co., Ltd. and its Yinchuan branch should pay rent to the defendant. In 2018, Cai obtained the right to use the shop and the rent claim from Song through the "Shop Right Transfer Agreement", and claimed to have notified the defendant. However, the defendant questioned Cai's subject qualification, arguing that the transfer agreement was not filed and no new authorization document was signed. Cai submitted relevant evidence to prove his rights, but the defendant raised objections to the validity of the evidence and demanded the deduction of taxes and litigation costs. The court needs to determine whether the defendant should bear the payment liability on the basis of clarifying the legality of the rent claim transfer and the effect of the notice.

Without Dispute Focus

This court has ascertained that on May 16, 2014, Song signed the "Entrustment Agreement for Leasing Commercial Properties" and the "Authorization Letter for Entrustment of Leasing Commercial Properties" with the defendant, Ningxia [redacted] Co., LTD. (hereinafter referred to as "Wurun Company") It is agreed that a total of 25 shops located at No. XXX, the second basement floor of [redacted] Commercial Plaza, South Gate, Yinchuan City, will be entrusted to Wurun Company for overall lease to Yinchuan Branch of Beijing [redacted] Co., LTD. (hereinafter referred to as "Beijing Hualian"), and the lease term, annual rent and payment method are clearly defined. After the agreement was signed, Wurun Company did not pay the rent to Song. On September 9, 2016, Wurun Company took Beijing Hualian to court, demanding payment of the lease fee. The court later ruled that Beijing Hualian should pay 17,328,990 yuan in rent to Wurun Company. On July 8, 2018, the plaintiff Cai signed a "Shop Usage Right Transfer Agreement" with Song, obtaining the usage rights of the above-mentioned 25 shops, and stipulated that Song would transfer his claims such as rental income against Wurun Company to Cai at the same time. Cai claimed that he had informed Wurun Company of the fact of the transfer, but Wurun Company failed to pay the rent. Wurun Company argued that the plaintiff Cai did not have the qualified subject status because Song did not file with Wurun Company and sign a new agreement and power of attorney when transferring the right to use the shop. Wurun Company acknowledges the "Entrustment Agreement for Leasing the Shop" signed with Song and is willing to pay the lease fee as per the court's judgment, but requests the deduction of taxes and litigation costs. During the trial, Cai submitted evidence such as the agreement on the transfer of the right to use the shop and the judgment of the Higher People's Court of Ningxia, attempting to prove that he had the right to demand that Wurun Company pay the rent. Wurun Company raised objections to the authenticity and legality of the evidence submitted by Cai, arguing that Cai failed to provide an effective notice of the transfer of the creditor's rights and did not sign a new agreement with Wurun Company. In conclusion, the crux of the dispute lies in whether the plaintiff Cai is a qualified subject

Figure 14: A failure case analysis for Fact generation in RQ2. The yellow background color indicates that this fact is mentioned both in the results with and without the dispute focus. The cyan background color indicates that this fact is only mentioned in the results without the dispute focus.

With Dispute Focus

Reasoning: Regarding the dispute focus: whether a private loan relationship between the plaintiff and the defendant is established. According to the Contract Law and relevant judicial interpretations … which is sufficient to prove that there is a loan relationship between the two parties. The defendant failed to provide evidence to prove that he had repaid the loan or that there was no loan relationship between the two parties. Therefore, the court determined that a private loan relationship was established between the plaintiff and the defendant. In summary, the court ordered the defendant Tao to return the principal of the loan of 120,000 yuan to the plaintiff Ma and pay the corresponding interest. The court did not support the plaintiff's other claims because they were not specifically listed.
Judgement Class: 2

Without Dispute Focus

Reasoning: The key points of this case are as follows: 1. The establishment of the loan relationship and the determination of the loan principal: The WeChat … sufficient to establish the existence of a loan relationship between the two parties 2. Calculation of interest: The plaintiff‘s claimed annual interest rate of … the plaintiff‘s claim for interest should be upheld. 3. Litigation costs: According to the … the defendant should bear the litigation costs of this case. In summary, the court supported the plaintiff Ma‘s lawsuit request and ordered the defendant Tao to repay the loan principal and pay the corresponding interest.
Judgement Class: 1

Ground Truth

Reasoning: The dispute in this case centers on whether a private loan relationship existed between the plaintiff and defendant. The underlying legal relationship in the dispute indicates a de facto partnership between the plaintiff, Ma, and the defendant, Tao. Both parties acknowledged the existence of this partnership in court. According to Article 16 of the Regulations, "The defendant shall provide evidence to support their claims. After the defendant has provided relevant evidence to support their claims, the plaintiff still bears the burden of proof for the existence of a loan relationship." The defendant argued in court that the funds involved were necessary for the operation of a company between the plaintiff, the defendant, and a third party, Zhai. This argument, consistent with the plaintiff's submission of telephone recordings, is highly probable and reasonable, and the defendant has fulfilled its burden of proof. The plaintiff asserted the existence of a loan relationship and should provide further evidence to prove a loan agreement between the plaintiff and defendant. However, the plaintiff failed to do so, and the evidence presented was insufficient to establish a loan agreement between the plaintiff, Ma Lei, and the defendant, Tao, for the funds involved. The fact that the plaintiff and defendant were in a partnership operating used cars does not rule out the possibility of a partnership debt between the parties involved. Prior to the judgment, the plaintiff failed to provide evidence, or the evidence was insufficient, to support its assertion. Therefore, the plaintiff's claim that the defendant should repay the loan of 120,000 yuan is insufficiently justified and has no legal basis, and this court does not support it.
Judgment Class: 0

Figure 15: A failure case analysis for Reasoning generation and LJP in RQ2. The yellow background color indicates the dispute focus. The cyan background color indicates the critical facts or articles of law when making a judgment.