Legal Fact Prediction: The Missing Piece in Legal Judgment Prediction

[1]Junkai Liu*, [3]Yujie Tong*, [1]Hui Huang, [1]Bowen Zheng, [4]Yiran Hu,
[3]Peicheng Wu, [2]Chuan Xiao, [2]Makoto Onizuka, [1]Muyun Yang†, [2]Shuyuan Zheng†
[1]Harbin Institute of Technology, [2]The University of Osaka,
[3]Zhejiang University, [4]Tsinghua University

Abstract

Legal judgment prediction (LJP), which enables litigants and their lawyers to forecast judgment outcomes and refine litigation strategies, has emerged as a crucial legal NLP task. Existing studies typically utilize legal facts, i.e., facts that have been established by evidence and determined by the judge, to predict the judgment. However, legal facts are often difficult to obtain in the early stages of litigation, significantly limiting the practical applicability of fact-based LJP. To address this limitation, we propose a novel legal NLP task: *legal fact prediction* (LFP), which takes the evidence submitted by litigants for trial as input to predict legal facts, thereby empowering fact-based LJP technologies to make predictions in the absence of ground-truth legal facts. We also propose the first benchmark dataset, LFPBench, for evaluating the LFP task. Our extensive experiments on LFPBench demonstrate the effectiveness of LFP-empowered LJP and highlight promising research directions for LFP.

1 Introduction

Advancements in NLP technology have significantly propelled the development of legal technology, particularly in the field of legal judgment prediction (LJP). LJP aims to predict court rulings based on litigation case information and legal provisions. For judges, automated predictions can serve as a reference for their official rulings, ensuring consistency in judicial standards. For litigants and their lawyers, pre- or in-trial judgment predictions help assess the potential outcomes of litigation, enabling them to make informed decisions. Consequently, LJP holds great potential for enhancing judicial efficiency and transparency.

Extensive research efforts have been devoted to achieving accurate LJP. However, existing LJP research is mostly limited to *fact-based LJP* (Luo et al., 2017; Zhong et al., 2018; Chen et al., 2019; Yue et al., 2021; Feng et al., 2022; Wu et al., 2022; Gan et al., 2023), where the input to the LJP system consists of (ground-truth) *legal facts*, i.e., facts that are formally established through evidence and determined by the judge. However, the users of LJP, such as litigants and lawyers, typically confirm their legal facts at a very late stage of the litigation (Medvedeva and Mcbride, 2023). Consequently, the application of LJP is largely confined, as the users often seek to predict judgments before litigation or in its early stages to develop and adjust litigation strategies or related plans.

To address the limitations of prior LJP studies, this paper proposes a novel legal NLP task: *legal fact prediction* (LFP), which aims to take the evidence submitted by litigants for trial as input and automatically determine relevant legal facts. Building on this foundation, we further introduce *LFP-empowered LJP*, as illustrated in Figure 1. In this approach, users first input available evidence into the LFP system to generate predicted legal facts, which are then used as the basis for the subsequent LJP task. This approach aligns more closely with real-world legal practice.

To further facilitate research on *LFP* and *LFP-empowered LJP*, this paper introduces the first benchmark dataset for LFP, *LFPBench*, which contains evidence items, legal facts, and judgment outcomes collected from 657 litigation cases in China, covering 10 representative types of civil cases. As such, it can be used to evaluate both the LFP and LJP tasks.

We conducted extensive experiments based on *LFPBench*, leveraging both general-domain and legal-domain models. The results reveal that, com-

*These authors contributed equally to this work and share first authorship.

†These authors share corresponding authorship: Muyun Yang <yangmuyun@hit.edu.cn>, Shuyuan Zheng <zheng@ist.osaka-u.ac.jp>.

Our code and data are available at https://github.com/teijyogen/LFP.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6334–6349
November 4-9, 2025

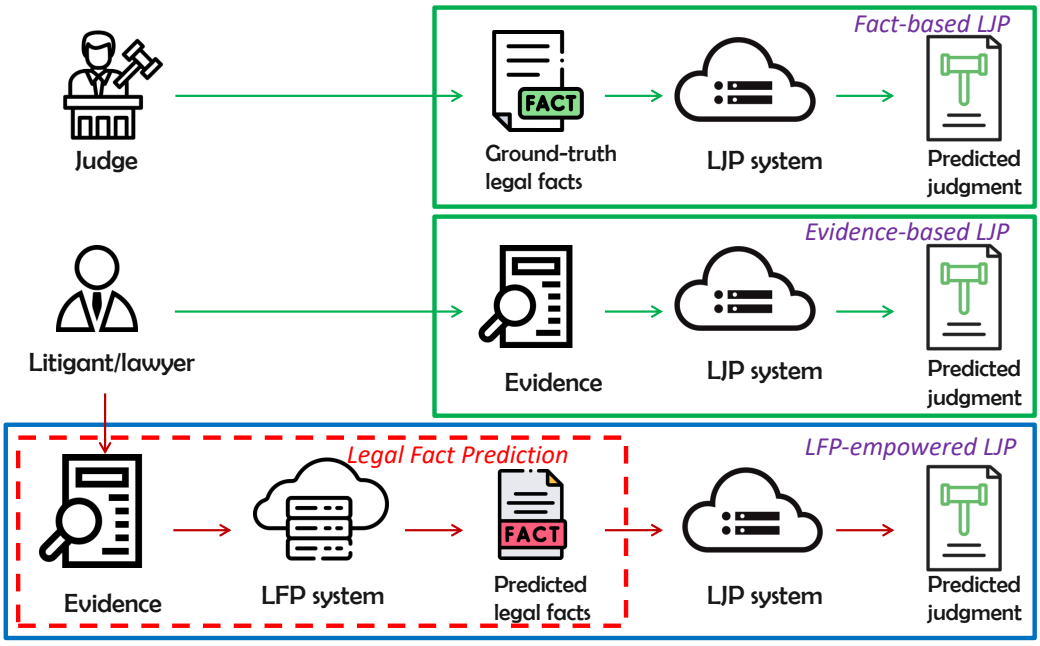

Figure 1: Connection between the legal fact prediction (LFP) and legal judgment prediction (LJP) tasks and comparison of three pipelines of LJP: *fact-based LJP*, *evidence-based LJP*, and *LFP-empowered LJP*. Most existing studies focus on fact-based LJP, while evidence-based LJP and LFP-empowered LJP remain unexplored.

pared to fact-based LJP, evidence-based LJP, where judgments are predicted solely based on evidence, exhibits a significant drop in accuracy. This suggests that the absence of legal facts has a profound impact on LJP. Moreover, LFP-empowered LJP reduces the accuracy drop of evidence-based LJP by 38.5% on average. Therefore, we argue that LFP is a crucial missing piece in the LJP task.

We summarize our contributions as follows:

- First, we propose a novel task, *legal fact prediction* (LFP), which empowers LJP applications to operate in a wider range of real-world scenarios.
- Second, we introduce *LFPBench*, the first benchmark dataset for studying LFP and LFP-empowered LJP, to support related research.
- Third, we conduct extensive experiments on LFPBench. Our results confirm the critical role of LFP in the LJP task and reveal the limitations of state-of-the-art (SOTA) models in addressing LFP. These findings offer valuable insights and guidance for future research.

2 The Legal Fact Prediction Task

In this section, we provide background information and formally define the LFP task.

2.1 Background

In the legal context, *evidence* refers to any material or information used to make the existence of a fact more or less probable (Wex, 2022a), whereas *legal facts*, also known as findings of fact, are the facts of a case determined by the judge during litigation, based on the presentation and cross-examination of evidence by the parties in a trial (Wex, 2022b). In other words, only facts that can be substantiated by evidence in a court of law can be acknowledged by the judge as legal facts.

As depicted in Figure 2, in civil law countries such as Germany, France, and China, as well as in common law countries like the UK and the US, a trial primarily resolves the following two tasks to reach a judgment:

- *Legal fact-finding*: Given the evidence presented and the arguments made by both the plaintiff and the defendant, the judge determines the legal facts of the case.
- *Application of law*: The judge applies the law to the legal facts to assess the validity of the plaintiff's claims and make an appropriate judgment.

As evident, legal facts serve as the foundation for the application of law, and before legal fact-finding is complete, it is logically impossible to predict a judgment based on legal facts. In fact,

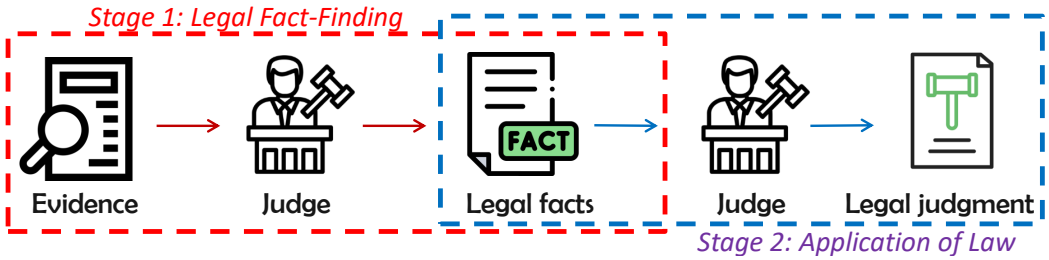

Figure 2: A trial primarily addresses two tasks: determining legal facts and applying the law.

Table 1: Comparison between LFPBench and existing LJP benchmarks.

Benchmark	Evidence items	Claims	Legal facts	Textual judgments	Judgment labels	Label classes	Type of cases	Jurisdiction
SwissJP (Niklaus et al., 2021)	✗	✓	✓	✓	✓	2	Generic	Switzerland
LJP-MSJudge (Ma et al., 2021)	✗	✓	✓	✗	✓	3	Civil	Mainland China
CAIL2018 (Xiao et al., 2018)	✗	✓	✓	✓	✓	2	Criminal	Mainland China
ILDC (Malik et al., 2021)	✗	✓	✓	✓	✓	2	Generic	India
Auto-Judge (Long et al., 2019)	✗	✓	✓	✗	✓	2	Civil	Mainland China
BrCase (Bertalan and Ruiz, 2020)	✗	✓	✓	✗	✓	2	Generic	Brazil
PhilCases (Virtucio et al., 2018)	✗	✓	✓	✗	✓	2	Criminal	Philippines
LFPBench (ours)	✓	✓	✓	✓	✓	3	Civil	Mainland China

legal facts are usually finalized when the judge reaches a judgment. Therefore, as Medvedeva and Mcbride (2023) has pointed out, utilizing ground-truth legal facts for LJP is impractical. Instead, litigants typically complete evidence collection before litigation or in its early stages, making *evidence-based LJP* a more reasonable choice. However, existing research on LJP primarily assume the accessibility of legal facts, namely they are all limited to *fact-based LJP* (e.g., (Luo et al., 2017; Zhong et al., 2018; Chen et al., 2019; Yue et al., 2021; Feng et al., 2022; Wu et al., 2022)), which is mismatched with real-world legal practice.

2.2 Task Definition

Next, we introduce the formal definition of the LFP task. Let C denote the plaintiff's claims, which determine the scope of the trial and constrain the space of legal facts to be predicted. Let Z denote the list of evidence for trial, which records all available evidence items to be presented and examined in establishing legal facts. Then, the LFP task requires a system f that takes the tuple (C, Z) as input and yields a set of legal facts $f(C, Z)$.

The predicted legal facts $f(C, Z)$, along with the claims C, can be further input into a given LJP system g to obtain the judgments $g(C, f(C, Z))$ for the claims. Since our motivation is to enhance evidence-based LJP, the objective of the LFP task is to find the optimal LFP system f that maximizes the accuracy of the predicted judgments $g(C, f(C, Z))$.

Note that the LFP task is not to summarize the evidence into legal facts. Instead, the available evidence items represent fragmented pieces of legal facts rather than a complete picture. Therefore, the LFP task involves deducing and expanding the evidence into legal facts. Moreover, conflicts or contradictions may exist among evidence items, particularly between the evidence presented by the plaintiff and the defendant. This requires the LFP system to assess the strength and logical coherence of the evidence to resolve the conflicts, making the prediction of legal facts a significant challenge. Notice the plaintiff's claims and the evidence list are typically available before the trial; therefore, leveraging this information as input for LFP aligns with real-world legal practice. We discuss the flexible adaptability of the input across different scenarios in Appendix A.

3 LFPBench

3.1 Dataset Overview

As shown in Table 1, existing LJP datasets, such as CAIL2018 (Xiao et al., 2018) and ILDC (Malik et al., 2021), focus on the prediction tasks of prison terms and charges rather than including the evidence items submitted by litigants, making LFP infeasible (Cui et al., 2023b). Therefore, we propose the first benchmark dataset for the LFP task, LFPBench. LFPBench consists of data for 657 first-instance cases in China, covering ten types of civil causes of action as shown in Figure 4. Each

Plaintiff's Claims	• **Claim 1:** Request the court to judge the defendant to pay the rent, property management fee, commercial promotion fee 45752.06 RMB, retroactive renovation period reduced rent, property management fee, commercial promotion fee and preferential period preferential rent 41846.4 RMB, termination of liquidated damages 44636.16 RMB, late payment of liquidated damages 9847.53 RMB, the total: 142082.15 RMB; • **Claim 2:** Request the court to judge the defendant bear all litigation expenses.	
Evidence Items	Submitted by Plaintiff	Submitted by Defendant
	• …… • **Evidence 4:** Lease Agreement, Request for Closure **Content:** Proving that:1) The Defendant leased the property from the Plaintiff for the operation of the restaurant; 2) After leasing the property, the Defendant paid only part of the rent and still owes a total of 142,082.15 RMB, which includes unpaid rent, property management fees, commercial promotion fees, outstanding payments for rent reductions, property management, commercial promotion, and discount-period rent; 3) Until July 7, 2023, the Defendant has not moved out.	• **……** • **Evidence 6**: Store Closure Application **Content:** Proving that the move-out date is May 31, 2023, and that the closure application has been submitted to the Plaintiff's office; • **Evidence 7:** The chat record between the Defendant and another Plaintiff's company manager. **Content:** Proving that the actual store closure date is May 31, 2023.
Legal Facts	The Defendant signed a "Lease Agreement" with the Plaintiff, leasing the Plaintiff's property for operating a restaurant…… On May 25, 2023, the Defendant sent the Plaintiff's company manager a "Store Closure Application" via WeChat……, which stated: "…… Formally submitting a move-out request to your company. The move-out date is May 31, 2023……" …… On July 7, 2023, the Plaintiff sent the "Request for Closure" to the Defendant, and the Defendant received it on July 14, 2023…… The Plaintiff now demands that the Defendant pay rent, property management fees, and commercial promotion fees for the period from January 1, 2023, to June 30, 2023.	
Judgements	• **Judgment for Claim 1:** The Defendant shall pay the Plaintiff rent, property management fees, and commercial promotion fees totaling 35,752.06 RMB within five days from the date this judgment becomes effective. • **Judgment for Claim 2:** The case acceptance fee is 1,571 RMB, of which 347 RMB shall be borne by the Defendant and 1,224 RMB by the Plaintiff.	
Labelled Judgments	• **Judgment for Claim 1:** partially supported • **Judgment for Claim 2:** partially supported	

Figure 3: A data sample from the LFPBench dataset featuring a house lease case. Both the plaintiff and the defendant submitted evidence to assist the judge in determining the legal facts. However, Evidence 4, Evidence 6, and Evidence 7 present conflicting information regarding the defendant's actual move-out date. Ultimately, according to Evidence 7, the judge determined that the defendant had not moved out before July 7 and had defaulted on the rent for June.

Table 2: Data statistics of the LFPBench dataset. Complete-win: all claims of the plaintiff are supported. Partial-win: part of the plaintiff's claims are completely or partially supported. Loss: all claims of the plaintiff are rejected.

No. of cases		No. of evid. items (plaintiff)	
Total	657	Max	19
With Defendant Evid.	387	Avg.	4.26
With Third-Party Evid.	80	Median	4
No. of cases results		**No. of evid. items (defendant)**	
Complete-win cases	166	Max	14
Partial-win cases	397	Avg.	1.83
Loss cases	94	Median	1
No. of judgment		**No. of claims**	
Full support	631	Max	9
Partial support	621	Avg.	2.48
Reject	378	Median	2

Figure 4: Distribution of case types in the LFPBench dataset.

case includes the plaintiff's claims, the evidence items submitted by the litigants, ground-truth legal facts, ground-truth judgments for the claims, and more. Therefore, LFPBench can be used for evaluating both LFP and LJP tasks. We have selected some widely used datasets for comparison with LFPBench, as shown in the Table 1. Not only does LFPBench include a third category for partial support by the court (the other two works only include support or opposition), but the input length also far exceeds theirs, posing a higher challenge to the model's capabilities. More importantly, to the best of our knowledge, LFPBench is the only dataset that considers predictions made before the trial in real-world scenarios, thus featuring a unique input of evidence lists, while other benchmark datasets follow the paradigm of using legal facts for judgment prediction.

The data statistics of LFPBench can be found in Table 2. In LFPBench, defendants present coun-

terevidence in 58.9% of cases, leading to disputes over the determination of legal facts. Consequently, only 38.71% of claims are fully supported, and plaintiffs completely win only 25.27% of the cases. Therefore, predicting the legal facts and judgments of these cases is highly challenging. Figure 3 presents a data sample from a house lease case, illustrating how conflicting evidence between the litigants complicates the determination of legal facts.

3.2 Dataset Construction

LFPBench data was extracted from judicial judgments in China. Our legal experts selected ten representative civil litigation causes of action with a moderate level of difficulty in establishing legal facts and retrieved 100 written judgments for each type from the China Judgments Online database (PRC, n.d.). These case types encompass various common disputes over property and personal rights in daily life. To ensure quality, three legal experts reviewed the judgments and excluded those with overly vague descriptions of evidence. Ultimately, 657 cases were retained.

Then, we used regular expressions to extract legal facts and judgment outcomes, as they are typically written in a consistent structure in the documents. Conversely, since the writing format for claims and evidence varies across judgments, we employed GPT-4o to extract this information. Afterward, our legal experts conducted a manual review to ensure consistency among the extracted claims, evidence, legal facts, and judgments.

Finally, our legal experts annotated the judgment outcomes. Specifically, as shown in Figure 3, they aligned each claim with its corresponding judgment and categorized the outcome into three labels based on the level of support: fully supported, partially supported, and rejected. These labels enable us to evaluate the LFP task using classification-based assessment methods.

3.3 Human Evaluation for Extracted Evidence

In many major jurisdictions such as China, Germany, France, and Japan, access to original evidentiary materials is, in principle, restricted to the parties involved in the case due to privacy considerations. Therefore, we opted to extract evidentiary information from publicly available judicial documents. Although such extracted evidence may lack some details compared to the original materials, the purpose and core content of the evidence are generally faithfully reflected in the judgment texts.

Table 3: Human evaluation of the quality of the extracted evidence in LFPBench, conducted by two legal experts. For comparison, the combined rates of affirmation and withdrawal in second-instance civil cases in China across different years are presented.

Metric	Value
Accuracy of Evidence-Based LJP by **Legal Experts**	87.62%
Affirmation + Withdrawl Rate in 2022	74.78%
Affirmation + Withdrawl Rate in 2023	75.61%
Affirmation + Withdrawl Rate in 2024	76.63%

Additionally, we asked our two legal experts to perform evidence-based LJP using our dataset, i.e., to predict the judgment outcomes solely based on the extracted evidence. They were required to evaluate a randomly selected sample of 100 cases, comprising a total of 259 plaintiff claims. As shown in Table 3, the experts achieved an LJP accuracy of 87.26%. This level of accuracy is notably high, considering that real-world judicial decisions are not entirely error-free. According to statistics released by the Supreme Court of China, the combined rates of affirmation and withdrawal in second-instance civil cases nationwide were 74.78%, 75.61%, and 76.63% in 2022, 2023, and 2024, respectively (PRC, 2025). These findings suggest that our extracted evidence retains the vast majority of critical information, enabling legal experts to make accurate judgments accordingly.

4 Experiment

4.1 Setup

Research questions. We conduct experiments to answer the following questions.

- RQ1 (Model & LJP Approach Comparison): How do SOTA models perform on the LFP and LJP tasks? How do different LJP approaches, including evidence-based LJP, fact-based LJP, and LFP-empowered LJP, perform?
- RQ2 (Challenge & Bias Analysis): What are the challenges of the LFP task? What biases do existing models exhibit when performing LFP?

Models. We employ 6 LLMs as the LFP and LJP systems, including the closed-source, general-purpose LLMs GPT-4o (OpenAI, 2024) and Claude-3.5-Sonnet-20241022 (Anthropic, Inc., 2025), the open-source, general-purpose LLMs

Table 4: Accuracy (%) of predicted judgments under different LJP approaches and models. **Def.**: cases where both parties have submitted evidence. **No def.**: cases where the defendant has not submitted evidence.

Model	Evidence-based			LFP-empowered			Fact-based		
	All	Def.	No def.	All	Def.	No def.	All	Def.	No def.
GPT-4o	50.67	50.91	50.31	**51.47**	49.39	54.67	55.77	52.02	61.53
Claude3.5	50.80	46.36	57.63	**52.58**	48.58	58.72	56.44	51.21	64.49
Qwen2.5-14B	45.09	42.21	49.53	**48.10**	43.83	54.67	49.45	44.74	56.70
Llama3.1-Chinese-8B	40.31	34.72	48.91	40.49	34.62	49.53	40.18	35.53	47.35
Average (General)	46.72	43.55	51.60	48.16	44.11	54.40	50.46	45.88	57.52
Law-Llama3.1-8B	31.10	28.85	35.36	30.12	26.72	40.65	33.13	32.89	33.49
LawJustice-Llama3.1-8B	35.21	30.97	41.74	28.96	26.42	32.87	32.33	26.42	41.43
Average (Legal)	33.16	29.91	38.55	29.54	26.57	36.76	32.73	29.66	37.46

Table 5: LFP similarities under different models.

	ROUGE	ChatLaw	LLM-as-Judge
GPT-4o	0.1808	0.7629	5.52
Claude3.5	0.2138	0.7668	5.83
Qwen2.5-14B	0.1692	0.7464	5.67
Llama3.1-Chinese-8B	0.1763	0.7549	6.23
LawLlama3.1-8B	0.1785	0.7455	5.46
LawJustice-8B	0.1721	0.7069	3.70

Qwen2.5-14B-Instruct (Yang et al., 2024) and Llama3.1-Chinese-8B (UnicomAI, 2024), as well as the open-source legal LLMs LawJustice-Llama3.1-8B (BAAI, 2024b) and Law-Llama3.1-8B (basuo, 2024). Additionally, we evaluated other legal LLMs including DISC-LawLLM (Yue et al., 2024), Lawyer-Llama-13B-V2 (Huang et al., 2023) and AIE-51-8-Law-Model (lingminai, 2025). However, these models failed to perform the LFP task due to poor instruction-following capabilities, as detailed in Appendix C.2.

LJP approaches. We compare the following LJP approaches that differ in their input, with their prompts detailed in Appendix C.4.

- *Evidence-based LJP*: The submitted evidence and the plaintiff's claims are directly input into the LJP system to generate legal judgments.
- *LFP-empowered LJP*: The submitted evidence and the plaintiff's claims are first input into the LFP system to predict legal facts. The predicted facts, along with the original inputs, are then fed into the LJP system to generate legal judgments.
- *Fact-based LJP*: The LJP system predicts legal judgments based on the ground-truth legal facts and the plaintiff's claims.

Metrics We primarily use **LJP accuracy** as the metric to quantify the influence of LFP on LJP. To measure **LFP similarity**—that is, the similarity between predicted and ground-truth legal facts—we employ three metrics: (1) ROUGE (Lin, 2004); (2) ChatLaw-based similarity, which calculates the distance between the predicted and ground-truth facts based on their embeddings determined by the ChatLaw-Text2Vecw model (Cui et al., 2023a); and (3) LLM-Judge scores assigned by GPT-4o on a ten-point scale. We also report F1 scores in Appendix C.3, which demonstrate consistency across metrics.

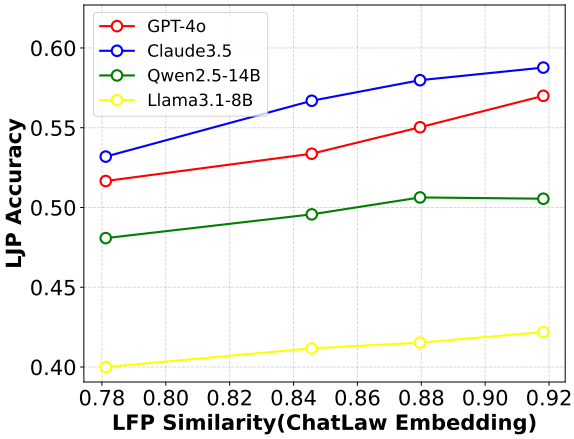

Figure 5: The correlation between the LFP similarity and the LJP accuaracy. We leverage the DP-Prompt method (Utpala et al., 2023) to generate rewritten legal facts with varying LFP similarities.

4.2 Model & LJP Approach Comparison (RQ1)

Finding 1: Legal LLMs perform poorly in LFP and LFP-empowered LJP. As shown in Table 4, the closed-source models GPT-4o and Claude3.5 consistently achieve the best performance across different LJP approaches, while the open-source legal LLMs Law-Llama3.1-8B and LawJustice-Llama3.1-8B perform the worst, with accuracy

Table 6: Accuracy (%) of LFP-empowered LJP for different judicial cases. **Def.**: cases where both parties have submitted evidence. **No def.**: cases where the defendant has not submitted evidence.

Model	Complete-win case			Partial-win case			Loss case		
	All	Def.	No def.	All	Def.	No def.	All	Def.	No def.
GPT-4o	66.95	31.06	88.29	44.49	47.85	38.02	60.47	71.52	29.82
Claude3.5	84.75	68.94	94.14	43.83	44.13	43.25	42.79	51.27	19.30
Qwen2.5-14B	75.99	60.61	85.14	45.15	45.99	43.53	16.74	20.25	7.02
Llama-3.1	75.42	59.09	85.14	34.31	34.38	34.16	13.49	15.19	8.77
Law-Llama3.1-8B	51.69	40.15	58.56	31.20	29.23	34.99	5.12	4.43	7.02
LawJustice-Llama3.1-8B	41.81	40.15	42.79	30.16	29.23	31.96	1.86	2.53	0.00
Average	**66.10**	**50.00**	**75.68**	**38.19**	**38.47**	**37.65**	**23.41**	**27.53**	**11.99**

Table 7: Accuracy (%) of LFP-empowered LJP for different causes of action. **LPR**: Labor Payment Recovery. **PC**: Pre-sale Contract. **SC**: Sales Contract. **ID**: Inheritance. **HL**: House Lease. **TL**: Tort Liability. **UE**: Unjust Enrichment. **PR**: Property Return. **MP**: Marital Property. **RLBH**: Right to Life/Body/Health.

Model	LPR	PC	SC	ID	HL	TL	UE	PR	MP	RLBH
GPT-4o	60.16	52.10	62.84	50.58	50.00	54.60	53.59	53.23	39.88	40.40
Claude3.5	66.41	61.08	63.51	55.81	54.49	45.98	51.63	51.08	38.15	41.72
Qwen2.5-14B	71.88	53.89	63.51	47.09	56.18	40.80	41.18	41.94	39.88	30.46
Llama3.1-Chinese-8B	63.28	45.51	52.03	45.35	39.33	34.48	32.03	33.87	31.21	34.44
Law-Llama3.1-8B	47.66	22.75	41.22	31.98	30.34	33.91	29.41	24.73	33.53	31.79
LawJustice-Llama3.1-8B	42.19	26.95	35.81	26.74	24.72	29.31	24.18	22.58	34.10	27.15
Average	**58.60**	**56.37**	**53.15**	**42.93**	**42.51**	**39.85**	**38.67**	**37.91**	**36.13**	**34.33**

close to random guessing. For the LFP performance in Table 5, the open-source legal LLMs again perform the worst, with the lowest performance among all metrics. This discrepancy cannot be attributed solely to the relatively small size of the legal LLMs, as Llama3.1-Chinese-8B, which has the same model size, performs significantly better. One possible explanation is that these legal LLMs are typically fine-tuned on short-text legal QA datasets (BAAI, 2024a; Yue et al., 2024; Huang et al., 2023), making them less capable of handling complex tasks such as LFP and LJP, which require summarization, reasoning, and deduction over long texts. Future research could explore incorporating general-domain instruction data into fine-tuning to mitigate the catastrophic forgetting of fundamental capabilities. It is also promising to develop more complex reasoning datasets in the legal domain.

Finding 2: Incorporating LFP can substantially reduce the performance gap between evidence-based LJP and fact-based LJP. As shown in Table 4, for the general-purpose LLMs, predicting legal judgments directly from evidence results in a $\frac{50.46-46.72}{50.46} = 7.42\%$ decrease in accuracy compared to predictions based on ground-truth legal facts. This indicates that while fact-based LJP research has made considerable progress, its effectiveness heavily relies on the accessibility of legal facts. On the other hand, although the more practice-aligned evidence-based LJP underperforms[1], the incorporation of LFP reduces the performance gap by 38.50%. Therefore, predicting legal facts from evidence first and then making legal judgments based on the predicted facts can significantly improve the accuracy of evidence-based LJP. Compared to fact-based LJP, LFP-empowered LJP accommodates a broader range of LJP scenarios in legal practice, striking a favorable balance between accuracy and applicability.

Finding 3: More accurate legal facts yield more accurate legal judgments. Using the DP-Prompt method Utpala et al. (2023), we rewrite the predicted legal facts with the Qwen2.5-14B-Instruct model, generating four versions with varying levels of LFP similarity. We then perform LJP on each version of the rewritten facts and repeated the experiment three times for each parameter setting. As shown in Figure 5, there is a positive correlation between the similarity of the rewritten legal facts and the accuracy of the corresponding legal judgments. These results further underscore the importance of the LFP task in enhancing LJP performance: more accurate legal facts lead to more accurate legal judgments.

[1]Note that for the legal LLMs, evidence-based LJP outperforms the other approaches. However, given their poor performance close to random guessing, this difference is likely due to randomness rather than a meaningful advantage.

Table 8: The proportions of cases in which each model predicts a complete win/partial win/loss for the plaintiff using LFP-empowered LJP. Evidence items are ordered, ensuring all evidence from the plaintiff (or defendant) appears first.

Model	Defendant first, plaintiff last				Plaintiff first, defendant last			
	Accuracy	Complete win rate	Loss rate	Partial win rate	Accuracy	Complete win rate	Loss rate	Partial win rate
GPT-4o	42.19	16.26	29.75	53.99	48.80	12.27	28.83	58.90
Claude3.5	39.78	42.33	1.84	55.83	49.64	37.42	18.40	44.17
Qwen2.5-14B	45.55	29.14	1.23	69.63	44.35	26.69	2.15	71.17
Llama3.1-Chinese-8B	35.23	19.69	2.77	77.54	34.74	22.09	2.45	75.46
Law-Llama3.1-8B	28.61	4.45	0.40	95.14	27.04	4.60	1.15	94.25
LawJustice-Llama3.1-8B	25.36	3.07	0.00	96.93	28.49	2.03	0.0	97.97
Average	**36.12**	**19.16**	**5.99**	**74.51**	**38.84**	**17.52**	**8.83**	**73.65**
Ground truth	100.00	17.18	17.18	65.64	100.00	17.18	17.18	65.64

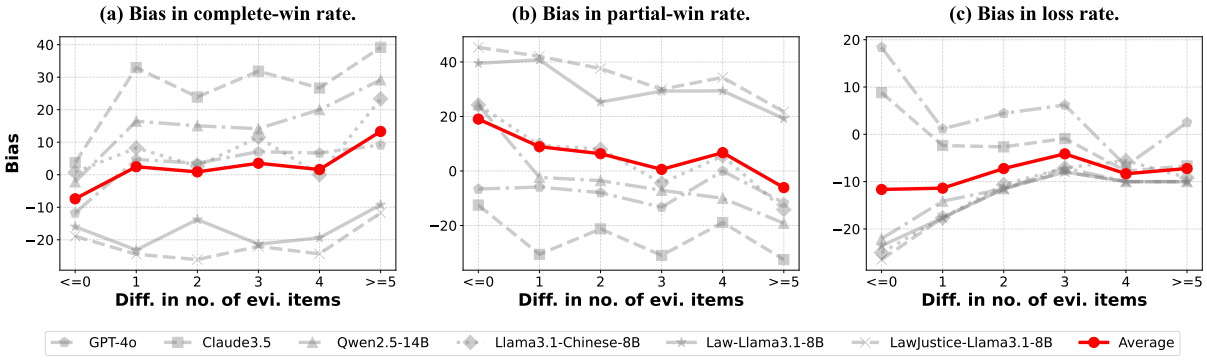

Figure 6: The effect of the difference in the number of evidence items between the plaintiff and the defendant on the judgments yielded by LFP-empowered LJP. The bias in y-axis means the difference in the rate between the model's predictions and the ground truth.

4.3 Challenge & Bias Analysis (RQ2)

Finding 4: Judgement prediction for partial-win or loss cases is more challenging. In Table 6, we report the performance of the LLMs in LFP-empowered LJP across cases where the plaintiff achieves a complete win, a partial win, or a loss. The results show that accuracy is typically highest in complete-win cases, while loss cases are the most challenging. This may be because, in real-world scenarios, complete-win cases usually have strong supporting evidence, making it easier to infer legal facts and judgments. In contrast, partial-win and loss cases often involve significant disputes between the parties, with conflicting or contradictory evidence, making it difficult to establish legal facts and reach a judgment.

Finding 5: Judgement prediction for cases with defendant evidence is more challenging. In Tables 4 and 6, we distinguish between cases where the defendant did or did not submit evidence. The results show that LFP-empowered LJP performs better in the former. This further suggests that counterevidence presented by the defendant can hinder LLMs' ability to infer legal facts. Therefore, future research should focus on enhancing LLMs' reasoning capabilities to better assess the authenticity of information.

Finding 6: Judgment prediction with weaker evidence is more challenging. From Table 7, we observe that cases involving LPR, PC, SC, ID, and HL, which typically feature strong written evidence (e.g., contracts and wills), allow for easier prediction of legal facts and judgments. Moreover, due to their textual nature, written evidence is more easily understood by LLMs. In contrast, cases such as TL and RLBH, which involve torts, often rely on non-written evidence, such as physical objects and audiovisual materials. When described in text, these forms of evidence lose significant detail. This suggests that future research could leverage multimodal technology to better interpret image- and sound-based evidence.

Finding 7: Bias arises from the presentation order of evidence items. In Table 8, we select all cases that include both plaintiff and defendant evidence, sort the evidence list in different orders, and then feed them to the models for LFP-empowered

LJP. We find that the order in which plaintiff and defendant evidence appears significantly influences the predicted judgments, introducing a bias favoring the party whose evidence is presented last. Specifically, when plaintiff evidence appears later, the LLMs are more likely to predict a complete or partial win for the plaintiff. Conversely, when defendant evidence appears later, the likelihood of the plaintiff losing the case increases. This bias may be attributed to the attention mechanism of LLMs (Yu et al., 2024), which requires further exploration in future research.

Finding 8: Bias arises from the number of evidence items. Figure 6 illustrates the impact of the difference in the number of evidence items between the plaintiff and the defendant on the predicted judgments. We observe that as the gap in evidence quantity increases, the LLMs generally tend to predict judgments with a higher complete-win rate and a lower partial-win rate compared to the ground truth. This suggests that an advantage in evidence quantity may lead LLMs to develop a bias toward fully supporting the plaintiff. However, the effect of this advantage on the loss rate is highly divergent: as the plaintiff's advantage increases, the two closed-source models become less likely to predict a loss for the plaintiff, whereas the open-source models exhibit the opposite. Nevertheless, more efforts are needed to teach LLMs that more evidence items don't necessarily mean stronger evidence or cause legal facts.

5 Related Work

Legal Fact Prediction Research on LJP can be traced back to the 1960s (Lawlor, 1963), which is one of the most fundamental tasks in legal AI. As judgment documents have become publicly accessible in many countries, researchers have extracted legal facts and judgment results from these documents, forming plenty of benchmark datasets for LJP research (e.g., (Xiao et al., 2018; Chalkidis et al., 2019; Malik et al., 2021; Semo et al., 2022; Chalkidis et al., 2022; Hwang et al., 2022)). Using judgment documents, numerous legal NLP studies have explored fact-based LJP, which predicts legal judgments based on legal facts, achieving promising predictive accuracy (e.g., (Luo et al., 2017; Zhong et al., 2018; Chen et al., 2019; Yue et al., 2021; Feng et al., 2022; Wu et al., 2022; Gan et al., 2023)). However, legal facts are not objective facts and are often difficult for LJP's intended users to obtain before a judgment is rendered. Consequently, Medvedeva et al. (Medvedeva and Mcbride, 2023) recently pointed out that most existing LJP studies rely on unrealistic input such as legal facts, limiting their practicality. Several studies employed legal briefs (Tippett et al., 2021), complaint documents (McConnell et al., 2021), court debate records (Ma et al., 2021) for LJP, but these types of judicial documents are typically not publicly accessible, making it impractical to obtain large-scale datasets for training LJP models. To address the lack of practicality in current LJP research, this paper proposes the LFP task as a preliminary step to fact-based LJP. LFP-empowered LJP establishes a practical loop from evidence to judgment, thereby making LJP more applicable in real-world scenarios.

Legal Document Summarization Legal Document Summarization (LDS) is the most correlated task with our LFP, which aims to automatically producing concise, accurate, and coherent summaries of legal texts (Kanapala et al., 2019), such as judgment documents (Polsley et al., 2016), contracts (Manor and Li, 2019), and court debate records (Duan et al., 2019). Intuitively, while LFP enriches and assembles fragmented and concise pieces of evidence into complete legal facts, LDS takes the opposite approach by refining legal texts and eliminating lengthy and complex details. Additionally, LFP requires inferring logically coherent factual information from conflicting or contradictory evidence, making it more challenging than text summarization.

6 Conclusion

This paper introduces the LFP task to automate the prediction of legal facts for the subsequent LJP task, addressing recent concerns that using ground-truth legal facts for LJP is impractical. We constructed a benchmark dataset for the LFP task, LFPBench, based on publicly available judicial documents. Extensive experiments conducted on LFPBench reveal that SOTA LLMs struggle to accurately determine legal facts when faced with conflicting or contradictory evidence, and exhibit biases related to the quantity and presentation order of evidence. Future work includes addressing the above limitations and constructing larger-scale LFP datasets to facilitate more extensive research.

Limitations

As the first step on the LFP task, this work has the following limitations. First, the evidence information in LFPBench is extracted from publicly accessible judgment documents in China, which are typically summarized and may lack some details of the original evidence, making it more challenging to predict legal facts. Second, the scope of case types covered by LFPBench remains limited, as it does not include criminal or administrative litigation cases. However, we would like to note that focusing on civil cases within a single jurisdiction is a common practice for benchmarking LJP (see Table 1).

Ethical Considerations

Our work may raise the following ethical considerations. (1) Data Privacy and Confidentiality: Judicial documents often contain some basic personal information of litigants, such as names, addresses, and identity numbers. Although we have processed the data to remove or anonymize personally identifiable information (PII), we must still comply with data usage regulations and refrain from any de-anonymization attempts that could compromise personal privacy. (2) Judicial Bias. Inappropriate applications of LJP may introduce ethical challenges, particularly because current fact-based LJP research often relies on legal facts extracted from court opinions, which may reflect judges or jurors' biases. As a result, such biases can be embedded in LJP's decisions. However, the introduction of the LFP task offers a way to alleviate this issue: it shifts the predictive foundation of LJP from biased legal facts in court opinions to facts predicted by LFP based on objective evidence. We believe this approach can, to some extent, reduce the influence of judicial bias. (3) Automated Adjudication. Some voices have proposed using LJP systems to replace human judges and juries, which has raised ethical concerns. However, we believe that the primary purpose of LFP and LJP is to assist litigants and their lawyers in predicting potential court opinions, thereby enabling them to adjust their strategies accordingly. This application can improve judicial transparency and reduce unnecessary judicial costs.

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

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62276077, 62376075, 62376076), JSPS Kakenhi (JP23K17456, JP23K25157, JP23K28096, JP25H01117, JP25K21207), and CREST (JPMJCR22M2).

References

Anthropic, Inc. 2025. Models overview — Claude.

BAAI. 2024a. IndustryInstruction_Law-Justice.

BAAI. 2024b. Law-Justice-llama3.1-8B-instruct.

basuo. 2024. llama-law.

Vithor Gomes Ferreira Bertalan and Evandro Eduardo Seron Ruiz. 2020. Predicting judicial outcomes in the Brazilian legal system using textual features. In *DHandNLP@ PROPOR*, pages 22–32.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-based prison term prediction with deep gating network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367.

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

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023b. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 11:102050–102071.

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*, pages 1361–1370.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664.

Leilei Gan, Baokui Li, Kun Kuang, Yating Zhang, Lei Wang, Anh Luu, Yi Yang, and Fei Wu. 2023. Exploiting contrastive learning and numerical evidence for confusing legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12174–12185.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer LlaMA. https://github.com/AndrewZhe/lawyer-llama.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for Korean legal language understanding and judgement prediction. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 32537–32551.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51:371–402.

Reed C Lawlor. 1963. What computers can do: Analysis and prediction of judicial decisions. *American Bar Association Journal*, pages 337–344.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. Workshop on Text Summariation Branches Out, Post-Conference Workshop of ACL 2004*, pages 74–81.

lingminai. 2025. AIE-51-8-Law-Model.

Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Automatic judgment prediction via legal reading comprehension. In *China National Conference on Chinese Computational Linguistics*, pages 558–572.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736.

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*, pages 993–1002.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4046–4062.

Laura Manor and Junyi Jessy Li. 2019. Plain english summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11.

Devin J McConnell, James Zhu, Sachin Pandya, and Derek Aguiar. 2021. Case-level prediction of motion outcomes in civil litigation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 99–108.

Masha Medvedeva and Pauline Mcbride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35.

OpenAI. 2024. Hello GPT-4o.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. CaseSummarizer: A system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, pages 258–262.

PRC. 2025. Judicial statistics in China.

PRC. n.d. China Judgments Online.

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31–46.

Elizabeth C Tippett, Charlotte S Alexander, Karl Branting, Paul Morawski, Carlos Balhana, Craig Pfeifer, and Sam Bayer. 2021. Does lawyering matter? Predicting judicial decisions from legal briefs, and what that means for access to justice. *Tex. L. Rev.*, 100:1157.

UnicomAI. 2024. Llama3.1-Chinese-8B-Instruct.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8442–8457.

Michael Benedict L Virtucio, Jeffrey A Aborot, John Kevin C Abonita, Roxanne S Avinante, Rother Jay B Copino, Michelle P Neverida, Vanesa O Osiana, Elmer C Peramo, Joanna G Syjuco, and Glenn Brian A Tan. 2018. Predicting decisions of the Philippine Supreme Court using natural language processing and machine learning. In *2018 IEEE 42nd Annual Computer Software and Applications Conference*, volume 2, pages 130–135. IEEE.

Wex. 2022a. Definition of evidence. Last reviewed in November of 2022.

Wex. 2022b. Definition of finding of fact. Last reviewed in December of 2022.

Yiquan Wu, Yifei Liu, Weiming Lu, Yating Zhang, Jun Feng, Changlong Sun, Fei Wu, and Kun Kuang. 2022. Towards interactivity and interpretability: A rationale-based legal judgment prediction framework. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4787–4799.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. 2024. Mitigate position bias in large language models via scaling a single dimension. In *First Workshop on Long-Context Foundation Models@ ICML 2024*.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. NeurJudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 973–982.

Shengbin Yue, Shujun Liu, Yuxuan Zhou, Chenchen Shen, Siyuan Wang, Yao Xiao, Bingxuan Li, Yun Song, Xiaoyu Shen, Wei Chen, et al. 2024. LawLLM: Intelligent legal system with legal reasoning and verifiable retrieval. In *International Conference on Database Systems for Advanced Applications*, pages 304–321.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549.

A Discussion

Although the input for LFP is defined as the evidence list and the plaintiff's claims, other trial-related information could also be incorporated into this task. As discussed by Medvedeva and Mcbride (2023), the ideal input for LJP should encompass any information available to the court or the parties at the time of performing LJP, such as complaints, defenses, and evidence submitted by the parties. This principle also applies to LFP. However, the information available to the court or the parties depends on the stage of the trial. For example, before filing a lawsuit, the plaintiff and the defendant may only have access to the evidence they personally possess. After filing, they gain access to each other's evidence and arguments regarding the legal facts. In this work, we chose the evidence list as the basic input for LFP because, at different stages of the trial, both parties have access to certain evidence.

Note that the "evidence list" here does not necessarily correspond to the final list of evidence submitted to the court, but rather represents the set of evidence items available to the parties at the current stage. Additionally, if the parties have access to other trial-related information, it can be incorporated as supplementary input to improve prediction accuracy. This suggests that in future work, we can adapt the LFP task to different trial stages by tailoring the input, thereby addressing various demands in legal practice.

B Additional Details on LFPBench

B.1 Annotation

Our dataset's annotators consist of four graduate students, two of whom have academic background in law, and two in computer science. They are all co-authors of this thesis, and therefore, no remuneration was provided. Before beginning the annotation process, the annotators unified the criteria for "supported", "partially supported", and "rejected". Generally speaking, if there is any conflict between the court's ruling and the content of a claim (such as a minor discrepancy in the amount of money), it cannot be considered that the court supports the claim. Correspondingly, if there is any overlap between the court's ruling and the content of a claim (such as the recognition of a small portion of the damages), it cannot be considered that the court rejects the claim.

B.2 Anonymization

Before extracting the relevant legal judgment information, we have already removed and replaced all sensitive information, such as the names and identification numbers of the litigants. Therefore, using this benchmark is safe and does not pose any risk of personal information leakage.

B.3 Copyright Issue

The case data used in LFPBench is sourced from the China Judgments Online (https://wenshu.court.gov.cn/), a website established by the Supreme People's Court of China for the publication of judgments issued by courts at all levels in China. According to Article 5 of the Copyright Law of the People's Republic of China: "This Law does not apply to: (1) laws, regulations, resolutions, decisions, orders, and other documents of a legislative, administrative, or judicial nature, and their official translations..." As such, the judgment data used in this paper falls under the exemption outlined in this provision and is not subject to the Copyright Law.

C Additional Details on Experiments

C.1 Documentation of LLMs

Table 9 shows the basic information of the LLMs involved in our paper. The Apache License 2.0 allows us to freely use these assets for academic research. One of the goals in training domain-specific models is to achieve performance comparable to larger models within specific domains using fewer parameters. Therefore, we have opted for a model with a relatively small parameter count, ranging between 8-14B. All these models possess Chinese language capabilities, making them suitable for our benchmark tests. We report the hyperparameters used for the LLMs in Table 10.

Table 9: Documentation of the used LLMs.

Artifacts	License	Parameter Scale	Language
Lawyer-Llama-13B-V2	Apache License 2.0	13B	Chinese
DISC-LawLLM	Apache License 2.0	13B	Chinese
Llama3.1-Chinese-8B	Apache License 2.0	8B	Chinese
LawJustice-Llama3.1-8B	Apache License 2.0	8B	Chinese/English
Law-Llama3.1-8B	Not specified	8B	Chinese
AIE-51-8-Law-Model	Not specified	3B	Chinese
Qwen2.5-14B	Apache License 2.0	14B	Multilingual
GPT-4o	Closed-source	Unknown	Multilingual
Claude3.5	Closed-source	Unknown	Multilingual

Table 10: Hyperparameter settings.

parameter	Close-source Models	Open-source Models
frequency_penalty	0	0
logprobs	false	false
presence_penalty	0	0
temperature	1	0.7
max_output_tokens	4,096	2048
top_p	1	0.8

C.2 Selection of Legal LLMs

We conducted a preliminary evaluation to select legal LLMs. Each model was tested on the entire LFPBench dataset to evaluate its performance on the LJP task, and we calculated the percentage of its outputs that could be successfully extracted by the evaluation script. If the model's results were recognized by the script, it indicates that the model could follow our instructions for LJP.

Table 11 reports the performance of the general-purpose LLM Llama3.1-Chinese-8B and five legal LLMs in the preliminary evaluation, including DISC-LawLLM (Yue et al., 2024) and Lawyer-Llama-13B-V2 (Huang et al., 2023), Law-Llama3.1-8B, LawJustice-Llama3.1-8B, and AIE-51-8-Law-Model (lingminai, 2025). The results show that while Llama3.1-Chinese-8B could fully follow our instructions, the long and complex instructions posed significant challenges for the legal LLMs. Finally, we selected the two legal LLMs with the highest success rate in instruction following for the main experiments in Section 4.

Table 11: Success rate (%) of different legal domain models in following our LJP instructions. For reference, the general-purpose open-source model Llama3.1-Chinese-8B fully complies with our instructions.

Models	Base Models	Instruction Following Rate (%)
Llama3.1-Chinese-8B	Llama3.1-8B (2024)	100.00
Law-Llama3.1-8B	Llama3.1-8B (2024)	88.74
LawJustice-Llama3.1-8B	Llama3.1-8B (2024)	80.06
DISC-LawLLM	Baichuan-13B (2023)	70.47
Lawyer-Llama-13B-V2	Llama2-13B (2023)	62.42
AIE-51-8-Law-Model	Qwen2.5-3B (2024)	51.09

C.3 F1 Scores

We report the macro-average and micro-average F1 scores for various models and methods in Table 12. These results align with the accuracy metrics presented in Table 4.

Table 12: The macro-f1 and micro-f1 metrics of the ternary classification under different LJP approaches and models.

Model	Evidence-based		LFP-empowered		Fact-based	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
GPT-4o	0.4957	0.5086	0.5073	0.5150	0.5503	0.5611
Claude3.5	0.4530	0.5083	0.4914	0.5264	0.5520	0.5780
Qwen2.5-14B	0.4109	0.4534	0.4335	0.4840	0.4725	0.5157
Llama3.1-Chinese-8B	0.3356	0.4101	0.3572	0.4125	0.3450	0.3971
Law-Llama3.1-8B	0.2825	0.3141	0.2859	0.3221	0.2523	0.2822
LawJustice-Llama3.1-8B	0.2714	0.3521	0.2289	0.2900	0.2422	0.3092

C.4 Prompt Templates

In this section, we present the prompt templates we used for evidence extraction, claim extraction, legal fact prediction, evidence-based LJP and fact-based LJP. We have performed initial adjustment to the prompt templates to ensure the performance of different models.

Prompt 1: Evidence Extraction

[Court Record]
the original text of the reference judgment paper
[Evidence List]
the reference evidence list
Please follow the format of the example above to extract a list of evidence from the provided trial records and output it in the form of a JSON list. Each element in the list should be a dictionary representing a piece of evidence, containing two key-value pairs: "Party Submitting Evidence" and "Content of Evidence." The Party Submitting Evidence should be one of [Plaintiff, Defendant, Third Party], while the Content of Evidence should be extracted directly from the original text of the trial records.
[Court Record]
the original text of the target judgment paper
[Evidence List]

Prompt 2: Claim Extraction

[Court Record]
the original text of the reference judgment paper
You need to extract all claims of plaintiff from the court record given above, and then organize them into such a list:
[Plaintiff's Claims]
[
*"**claim1**",*
*"**claim2**",*
...
]
Each claim in the list should be as faithful to the original text as possible. Focus on the subjective opinions put forward by the defendant, and do not pay attention to his specific evidence. You only need to output the formatted list of claims, without adding any comments.
[Plaintiff's Claims]

Prompt 3: Legal Fact Prediction

[Plaintiff's Claims]
*(1) **claim1***
*(2) **claim2***
...
[Litigant]
the parties concerned
[Evidence List]
*(1) **submitting party***
content
*(2) **submitting party***
content
...
Please analyze the plaintiff's claims and the list of evidence in the above case, and output a faithful description of the basic facts of the case from the court's perspective.
Only provide the findings of fact, without adding any reasoning process or explanations.

Prompt 4: Evidence-Based LJP

[Case Type]
one of the ten types
[Litigant]
the parties concerned
[Evidence List]
*(1) **submitting party***
content
*(2) **submitting party***
content
...
[Plaintiff's Claims]
*(1) **claim1***
*(2) **claim2***
...
You need to refer to the evidence presented by all parties in the [Evidence List] to predict the court's judgment on the [Plaintiff's Claims], and form a corresponding judgment list.
The [Judgment List] is a list composed of three numbers (0, 1, -1). If you believe the court will fully support the claim, the result is 1; if partially supported, the result is 0; if the claim is dismissed, fill in -1.
Just output the formatted judgment list without any comments.

Prompt 5: Fact-Based LJP

[Case Type]
one of the ten types
[Litigant]
the parties concerned
[Reference Facts]
the fact determined by court
[Plaintiff's Claims]
*(1) **claim1***
*(2) **claim2***
...
You need to refer to the evidence presented by all parties in the [Evidence List] and the reference facts provided by the model in the [Reference Facts] to predict the court's judgment on the [Plaintiff's Claims] and form a corresponding judgment list.
The [Judgment List] is a list composed of three numbers (0, 1, -1). If you believe the court will fully support the claim, the result is 1; if partially supported, the result is 0; if the claim is dismissed, fill in -1.
Just output the formatted judgment list without any comments.