Beyond Guilt: Legal Judgment Prediction with Trichotomous Reasoning

Anonymous ACL submission

Abstract

In legal practice, judges apply the *trichotomous dogmatics of criminal law*[1], sequentially assessing *the elements of the offense*, *unlawfulness*, and *culpability*[2] to determine whether an individual's conduct constitutes a crime. Although current legal large language models (LLMs) show promising accuracy in judgment prediction, they lack trichotomous reasoning capabilities due to the absence of an appropriate benchmark dataset, preventing them from predicting "innocent" [3] outcomes. As a result, every input is automatically assigned a charge, limiting their practical utility in legal contexts. To bridge this gap, we introduce LJPIV, the first benchmark dataset for Legal Judgment Prediction with Innocent Verdicts. Adhering to the trichotomous dogmatics, we extend three widely-used legal datasets through LLM-based augmentation and manual verification. Our experiments with state-of-the-art legal LLMs and novel strategies that integrate trichotomous reasoning into zero-shot prompting and fine-tuning reveal: (1) current legal LLMs have significant room for improvement, with even the best models achieving an F1 score of less than 0.3 on LJPIV; and (2) our strategies notably enhance both in-domain and cross-domain judgment prediction accuracy, especially for cases resulting in an innocent verdict.

1 Introduction

The trichotomous dogmatics (Elias, 2015; Dubber, 2005) is a key theory in criminal law used to determine whether an individual's conduct constitutes a crime. It is widely applied in civil law jurisdictions such as Germany, Japan, and China. In legal practice, judges apply this framework in three sequential steps: (1) the elements of the offense, where the individual's conduct is assessed to see if it objectively fulfills the criteria for a criminal offense; (2) unlawfulness, where it is determined whether the individual has any grounds for justification—such as self-defense—that would exempt them from criminal liability despite fulfilling the offense's elements; and (3) culpability, which requires that the individual acted with free will to be held responsible. If the individual acted while mentally ill or lacking judgment, they are deemed to lack criminal responsibility and thus cannot be considered guilty. As illustrated in Figure 1, through trichotomous reasoning, the conduct must satisfy the elements of the offense, unlawfulness, and culpability in order to constitute a crime.

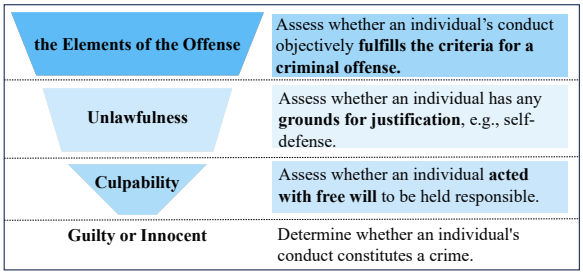


Figure 1: An illustration of trichotomous dogmatics of criminal law. The reasoning process proceeds sequentially from top to bottom. Conduct that satisfies the elements of the offense, unlawfulness, and culpability is deemed to constitute a crime; otherwise, the individual is considered innocent.

In the field of legal NLP (Wu et al., 2023; Cui et al., 2023; Shu et al., 2024), researchers fine-tune large language models (LLMs) on large-scale legal judgment prediction (LJP) corpora to enable them to predict applicable criminal charges based on an input fact description. While these models (Yue et al., 2023; Wu et al., 2023) have achieved impressive accuracy in charge prediction, they only map an individual's conduct to the elements of a criminal offense—the first step in trichotomous dogmat-

[1]Translated from the German legal term "der trichotomisch aufgebauten Dogmatik des Strafrechts".

[2]Translated from the German legal terms "Tatbestand", "Rechtswidrigkeit", and "Schuld".

[3]In this paper, the terms "non-guilty" and "innocent" are used interchangeably.

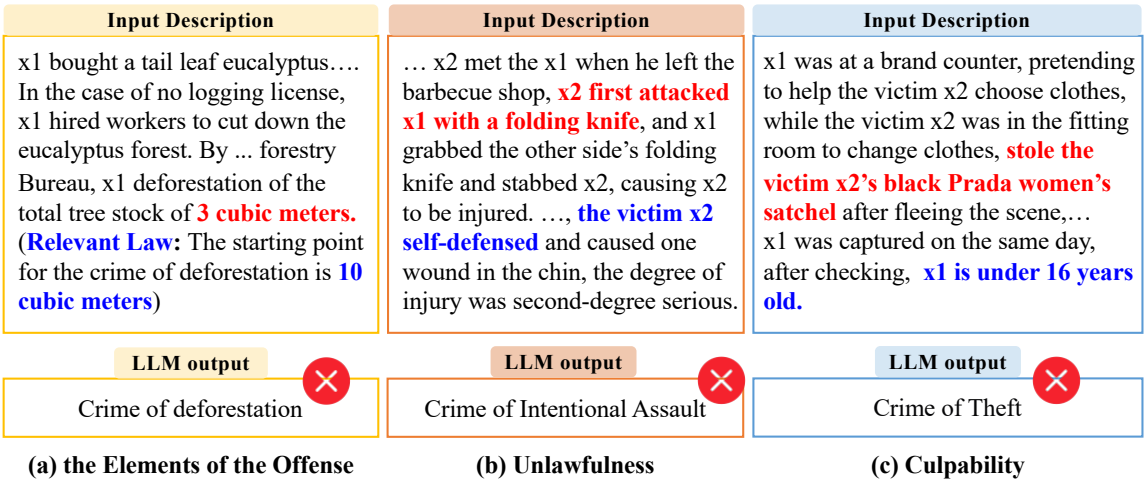


Figure 2: DISC-Law (Yue et al., 2023) incorrectly predicts charges for non-guilty fact descriptions across the elements of the offense, unlawfulness, and culpability. The red parts represent actions that may lead to a guilty verdict, while the blue parts indicate acts or situations that result in contradictions or exoneration.

ics—and thus lack the full trichotomous reasoning capability. As a result, a major limitation of existing legal LLMs is that they automatically assign a charge to the input without the ability to predict an "innocent" outcome, even when there are clear grounds for justification. For instance, as shown in Figure 2, DISC-Law (Yue et al., 2023), the state-of-the-art Chinese legal LLM, overlooks self-defense in example 2(b) and an age requirement for criminal responsibility in 2(c), leading to incorrect crime predictions. This limitation significantly reduces their practical utility in legal contexts. The primary cause of this issue is the lack of a benchmark dataset designed to support trichotomous reasoning in legal judgment prediction.

To address this gap, we introduce LJPIV, the first benchmark dataset for Legal Judgment Prediction with Innocent Verdicts. Unlike previous benchmarks that only provide fact descriptions and corresponding charge labels, LJPIV focuses on trichotomous reasoning, particularly for cases with innocent outcomes, by offering fine-grained, sentence-level labels that indicate compliance with the trichotomous dogmatics of criminal law. To construct LJPIV, we extend three popular benchmarks—CAIL (Xiao et al., 2018; Shui et al., 2023), ELAM (Yu et al., 2022b), and LeCaRD (Ma et al., 2021; Deng et al., 2024)—through a three-stage augmentation process. *First*, recognizing that some sentences in the fact descriptions (such as suspect profiles) may be less informative for judicial decisions, we fine-tune an LLM to extract sentences relevant to trichotomous reasoning, reducing the risk of over-correction (Li et al., 2023; Fang et al., 2023) in the subsequent stages. *Second*, we apply the retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023) technique to retrieve relevant criminal laws for the input and prompt the LLM to introduce grounds for justification—legal reasons that exempt the individual from criminal liability despite their conduct meeting the elements of the offense. This allows us to randomly select a portion of the data and create counterfactual samples[4] labeled as "innocent". *Third*, to ensure the logical consistency and coherence of these counterfactual samples, we prompt the LLM to perform a self-check, followed by multiple rounds of manual verification by annotators.

We further introduce zero-shot prompt-based and fine-tuning methods to equip open-domain LLMs (Bai et al., 2023) with trichotomous reasoning capabilities, particularly for predicting innocent outcomes in legal judgment prediction. Extensive experiments on state-of-the-art legal LLMs reveal that (1) current legal LLMs have significant room for improvement, with even the best models achieving an F1 score below 0.3 on LJPIV; (2) fine-tuning on LJPIV substantially improves both in-domain and cross-domain judgment prediction accuracy for open-domain LLMs, especially in cases resulting in an innocent verdict; and (3) our trichotomous reasoning strategies including the prompt-based and fine-tuning approaches further enhance the legal judgment performance.

[4]Please note that current LJP datasets are derived from publicly available cases, all of which involve guilty verdicts. Therefore, we refer to augmented samples corresponding to not-guilty situations as counterfactual samples.

Dataset	LLM Annotation	Non-guilt Label	Trichotomous Reasoning
CAIL-2018 (Xiao et al., 2018)	✗	✗	✗
CAIL-Long (Xiao et al., 2021)	✗	✗	✗
ELAM (Yu et al., 2022b)	✗	✗	✗
LeCaRD (Ma et al., 2021)	✗	✗	✗
DPAM (Wang et al., 2018)	✗	✗	✗
SLJA (Deng et al., 2023)	✓	✗	✗
LJPIV (Ours)	✓	✓	✓

Table 1: Comparison between existing LJP datasets with our LJPIV. LJPIV is the only one that includes not-guilty labels and supports trichotomous reasoning.

Our contributions can be summarized as follows:

- We identify a significant limitation of current legal LLMs: their inability to predict "Innocent" outcomes for given fact descriptions, which restricts their practical utility in legal contexts. Inspired by the trichotomous dogmatics of criminal law, our work pioneers the integration of trichotomous reasoning capabilities into LLMs for legal judgment prediction, particularly for predicting innocent outcomes.
- We construct the first benchmark dataset for legal judgment prediction with innocent verdicts by extending three popular benchmarks through a three-stage LLM-based augmentation process, followed by manual verification.
- Extensive experiments demonstrate that fine-tuning on LJPIV significantly enhances both in-domain and cross-domain judgment prediction accuracy for open-domain LLMs, particularly for cases with innocent verdicts, while also validating the effectiveness of our trichotomous reasoning strategies in further improving judgment performance.

2 Related Work

2.1 Legal Judgment Prediction

Legal judgment prediction is a classic legal task, aiming to predict a charge based on the case facts. We review the datasets for legal judgment prediction, as summarized in Table 1. The CAIL-2018 (Xiao et al., 2018) initiated the task of Chinese legal judgment prediction. ELAM (Yu et al., 2022b) and LeCaRD (Ma et al., 2021), have been proposed for judicial case matching tasks, where each case includes a corresponding case description and charge. These datasets have been used in legal judgment prediction tasks (Sun et al., 2024; Qin et al., 2024). These datasets primarily rely on manual annotation, requiring the hiring of professional legal workers and consuming significant time and resources. SLJA (Deng et al., 2023) is a legal judgment prediction dataset derived using syllogistic reasoning. Differing from these datasets, this paper introduces datasets built around a smaller-scale LLM using trichotomous reasoning that includes not-guilty cases.

2.2 LLM in Legal NLP

With the rapid development of LLMs, their performance in various open-domain tasks (Peng et al., 2023; Achiam et al., 2023) has been remarkable. The exploration of LLM applications in legal NLP tasks is burgeoning (Fei et al., 2023; Choi et al., 2021). By incorporating legal reasoning steps into the prompts (Blair-Stanek et al., 2023; Yu et al., 2022a) of LLMs, these models are guided to complete specific legal tasks. Additionally, techniques such as Retriever-Augmented Generation (RAG) (Zhang et al., 2023; Zhou et al., 2023; Pipitone and Alami, 2024) are utilized to retrieve relevant case law and content from legal knowledge bases to assist LLMs in completing legal tasks. Furthermore, by fine-tuning open-domain LLMs with extensive legal documents and legal task datasets, a surge in specialized legal LLMs (Yue et al., 2023; Cui et al., 2023; Shu et al., 2024), has been observed. The objective of this paper is to utilize LLMs to annotate a dataset that includes not-guilty cases and to explore how LLMs can implement trichotomous reasoning.

3 Dataset Construction

We focus on trichotomous reasoning for the legal judgment prediction task, which can be formulated as follows: given an input fact description x, the goal is to predict its judgment outcome $y \in \mathcal{Y}$. Unlike existing studies where $\mathcal{Y}$ represents a set of criminal charges, we consider a more practical scenario by extending $\mathcal{Y}$ to include a label for "innocent". Specifically, we utilize LLMs to augment three popular LJP datasets and create a new benchmark, LJPIV, through a three-stage augmentation.

3.1 Sentence Extraction for Reasoning

When utilizing LLMs to extend legal datasets, over-modification (Li et al., 2023; Fang et al., 2023) must be carefully considered, as it is unacceptable for legal documents due to their serious nature; excessive modifications can alter the legal meaning of the text and negatively impact trichotomous reasoning. Additionally, some sentences, such as

suspect profiles, may be less informative for the judicial decisions. Therefore, in the first stage, we extract sentences that correspond specifically to trichotomous reasoning.

To achieve this goal, we construct an instruction fine-tuning dataset, D_E, using LeCaRD-Elem (Deng et al., 2024), where the queries contain legal element annotations. Based on D_E, we develop an instruction template in the format $\{Instruction, (Case, Crime), Crime-related\ Sentence\}$ and use LoRA (Hu et al., 2021) to efficiently fine-tune LLM (Bai et al., 2023). This process allows us to inject the necessary knowledge for extracting reasoning-related sentences into the LLM, resulting in the sentence extractor E_C.

To extend existing LJP datasets, we segment them into sentences and apply E_C to extract those relevant to trichotomous reasoning. Given the relatively small size of D_E (approximately 100 entries), we also implement a decoding constraint (Geng et al., 2023; Lu et al., 2024) on E_C that adjusts the LLM's probability distribution during decoding to ensure each sentence is extracted only once.

3.2 Injection of Grounds for Justification

After extracting sentences relevant to trichotomous reasoning from the input fact descriptions, we randomly select 50%[5] (Considering the actual situation[6], it can be used for pre-court analysis and legal aid, and then relieve the pressure of the court) samples from each dataset and leverage the LLM's strong semantic understanding and instruction-following capabilities (Ouyang et al., 2022; Achiam et al., 2023) to generate counterfactual samples, guiding the LLM to inject grounds for justification into the extracted sentences. We use carefully designed prompts as shown in Figure 3(a), within a RAG framework (Lewis et al., 2020; Gao et al., 2023) to introduce acts or situations that result in contradictions or exoneration at the levels of the elements of the offense, unlawfulness, and culpability, based on the retrieved criminal laws relevant to the case.

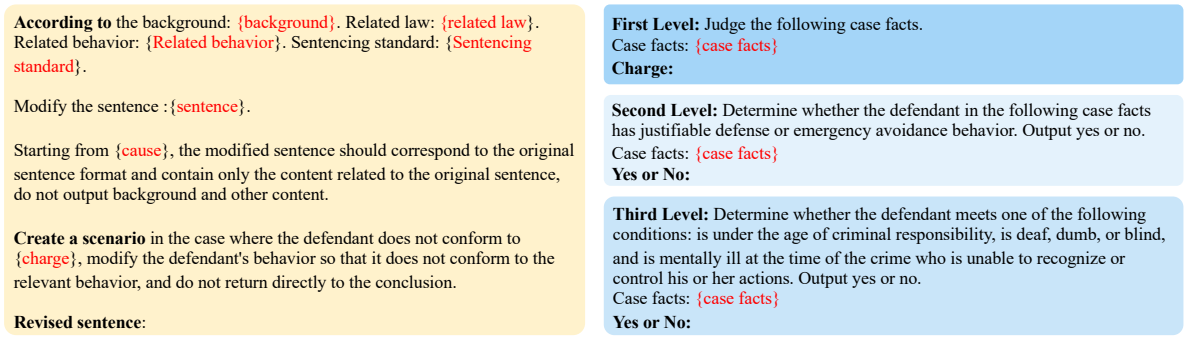
At the level of **the Elements of the Offense**, judges assess whether an individual's conduct objectively fulfills the criteria for a criminal offense. Considering that the current LJP datasets consist of already adjudicated cases, where each sample has assigned labels such as applicable charges and relevant legal articles, we derive exonerating actions and circumstances based on the legal definitions of the charges to construct counterfactual samples. Specifically, we use the charge as a query to retrieve the corresponding legal articles, criminal behaviors, and judgment criteria. Using these retrieval results, we guide the LLM to generate scenarios or actions that contradict the establishment of the charge and incorporate them into the original fact description to create not-guilty samples. For example, as shown in Figure 2(a), the legal definition for the crime of deforestation requires that the total volume of trees felled reaches 10 cubic meters. We instruct the LLM to modify the total volume of trees felled in the original case to below 10 cubic meters, such as 3 cubic meters, making the charge invalid.

At the level of **Unlawfulness**, judges assess whether an individual has any grounds for justification, meaning legal reasons that exempt a person from criminal liability despite fulfilling the elements of the offense. In Chinese criminal law, two primary situations lead to exoneration at the level of unlawfulness: self-defense and necessity. Therefore, we instruct the LLM to modify the original fact description by introducing a scenario involving either self-defense or necessity, so that the case satisfies the exoneration conditions for the charge. For example, as shown in Figure 2(b), although x2 caused second-degree harm to x1, his actions fall under the category of self-defense, meaning that x2 is not guilty of intentional injury. Please note that not all charges are applicable to self-defense or necessity. For instance, in the deforestation case shown in Figure 2(a), so we skip this step for such fact descriptions.

At the level of **Culpability**, judges assess whether an individual acted with free will and can be held responsible. In this study, we primarily consider three situations: (1) the defendant has not yet reached the age of criminal responsibility, (2) the defendant is deaf, mute, or blind, or (3) the defendant is a person with a mental illness who was unable to recognize or control their actions at the time of the crime. Since these situations are related to the description of the defendant, we randomly select one of these scenarios and instruct the LLM to incorporate it into the defendant's profile. For example, as shown in Figure 2(c).

[5]Users can sample a smaller percentage of "not guilty" cases (e.g., 10%) and combine them with "guilty" cases to construct a sub-dataset tailored to their needs.

[6]https://www.spp.gov.cn/xwfbh/wsfbt/202403/t20240310_648482.shtml



(a) Prompt for Data Construction **(b) Prompt for LLM Inference**

Figure 3: The prompt for trichotomous reasoning used in this study.

3.3 Data Quality Verification

Through manual inspection, we find that LLM-based augmentation can lead to over-modification. For example, the LLM may incorrectly treat a fracture as a minor injury, which contradicts common sense, and should instead modify it to a truly minor injury, such as a scratch. To address this logical inconsistency, we implement both an LLM self-check and manual verification to ensure the quality of our LJPIV dataset.

For the LLM self-check, we instruct LLM to check whether augmented cases, i.e. those counterfactual samples with non-guilty labels, have the aforementioned issues with the prompt *"Determine whether the following case facts have logical problems, common sense errors, contradictions, unreasonable or incoherent content"*.

After the LLM's self-check, we employ five legal annotators to manually verify the correctness of the augmented samples from both legal and logical perspectives. Specifically, we use a multi-round random inspection process. In each round, each annotator randomly selects 20% of the augmented samples and compares the fact descriptions against the legal provisions to ensure the samples are consistent with legal innocence. After each round, five annotators collaboratively revise the over-modified samples based on legal standards. The revised ones are excluded from the next round of inspection. This process is repeated until no issues are found in the randomly selected cases. Ultimately, we conducted five rounds of inspection and corrections. More information is in the Appendix A.

4 Trichotomous Reasoning with LLM

In this section, we introduce both prompt-based and fine-tuning methods to enable LLMs to perform trichotomous reasoning for LJP.

4.1 Prompt-Based Method

Given an input fact description x, with our carefully designed trichotomous prompts p_1, p_2, and p_3, as illustrated in Figure 3 (b), the LLM generates three predictions y_1, y_2, and y_3. Each prediction corresponds to the reasoning outcome for the levels of the elements of the offense, unlawfulness, and culpability, respectively:

$$y_k = f_{\text{LLM}}(x, p_k; \theta), \tag{1}$$

where θ represents the parameters of the LLM; $k \in \{1, 2, 3\}$ denotes the index for the reasoning level. Specifically, y_1 represents a charge or non-guilt prediction, while y_2 and y_3 are "Yes/No" responses from the LLM, indicating whether there are grounds for justification and whether the individual has criminal responsibility, respectively. Therefore, the overall judgment prediction y_{final} can be derived as follows:

$$y_{\text{final}} = \begin{cases} y_1 & \text{if } y_2 = \text{"No"}, y_3 = \text{"No"}, \\ \text{non-guilt} & \text{otherwise} \end{cases} \tag{2}$$

4.2 Fine-Tuning-Based Method

Fine-tuning the LLM using our LJPIV in the following two steps is an alternative approach to equip the LLM with trichotomous reasoning capabilities.

Fine-tuning Dataset D_{NG} Construction. For each level of trichotomous reasoning, the input for fine-tuning consists of the fact description and the corresponding prompt, as shown in Figure 3(b). In terms of the output, at the level of the elements of the offense, the output is either a criminal charge or an innocent label. At the levels of unlawfulness and culpability, the output is a "Yes/No" response, indicating whether there are grounds for justification and whether the individual has criminal responsibility, respectively.

Fine-tuing and Inference. We perform fine-tuning using LoRA (Hu et al., 2021) on D_{NG} :

$$\mathcal{L}_{\mathrm{FT}} = -\frac{1}{|\mathcal{D}_{NG}|} \sum_{\mathcal{D}_{NG}} \log(P_{\theta+\theta_L}(y_t|x, p, y_{<t})), \quad (3)$$

where θ and θ_L represent the parameters of the LLM and LoRA, respectively; y_t denotes the t-th token, and $y_{<t}$ represents the tokens preceding y_t. In the inference phase, we utilize the fine-tuned LLM to make the judgment in three sequential steps:

$$y_k = f_{\mathrm{LLM}}(x, p_k; \theta + \theta_L), \quad (4)$$

where $k \in \{1, 2, 3\}$ denotes the index for the reasoning level in trichotomous dogmatics.

5 Experiments

5.1 Experimental Settings

5.1.1 Dataset and Metric

Dataset. We conduct extensive experiments on the proposed LJPIV datasets, which are constructed by extending CAIL-2018 (Xiao et al., 2018), ELAM (Yu et al., 2022b), and LeCaRD (Ma et al., 2021), referred to as LJPIV-CAIL, LJPIV-ELAM, and LJPIV-LeCaRD, respectively. Each sample in LJPIV consists of a fact description and the corresponding judgment (non-guilty or a charge). The ratio of guilty to innocent samples is set to 1:1. Among the innocent samples, the reasons for innocence, corresponding to the three levels of trichotomous dogmatics, are balanced with a ratio of 3:1:1. Specific statistical information about these datasets is provided in Table 2.

Since both LLM-based and manual annotating are costly, we follow (Shui et al., 2023) and select a subset of CAIL-2018, called CAIL-train, as the training set for LJPIV. The remainder of CAIL-2018, along with ELAM and LeCaRD, serve as in-domain and cross-domain test sets for LJPIV. In addition, since ELAM and LeCaRD are based on retrieved datasets where only the queries have corresponding crime labels, we limit the test set to include only queries to ensure the quality of the judgment labels. Annotating crime labels for the candidate cases is reserved for future work.

Metric. Following (Feng et al., 2022; Zhong et al., 2018), we evaluate the LJP results using widely-used metrics including Accuracy (Acc), Precision (P), Recall (R), and F_1-Score (F_1).

Dataset	#Case	#Charge	Avg_Case_Len	Avg_Sent_Num
LJPIV-CAIL-Train	1120	112	439.43	6.23
LJPIV-CAIL-Test	560	112	436.63	6.16
LJPIV-ELAM	500	63	832.82	10.50
LJPIV-LeCaRD	80	30	448.34	7.25

Table 2: Statistics of our LJPIV. **#Case**, **#Charge** denote the number of the cases and charges. **Avg_Case_Len** and **Avg_Num_Sent** represent the average length of the case and the average number of sentences.

5.1.2 Baseline

(1) **Legal LLMs**: We selected three popular legal LLMs for comparison: **DISC-LawLLM** (Yue et al., 2023), **LexiLaw**[7], and **fuzi.mingcha** (Wu et al., 2023). These models are fine-tuned on a large amount of legal NLP task data, including legal judgment prediction datasets.

(2) **Open-domain LLMs**: We select Qwen (Bai et al., 2023) and Baichuan (Yang et al., 2023) as open-domain LLMs for our experiments, which have strong performers across various fields. We tested their judgment prediction abilities using the following methods: **Zero-shot**: We directly asked the LLMs to predict convictions. **Zero-shot-CoT**: While asking the LLMs to predict, we specifically instructed them to pay attention to innocence. **Few-shot-BM25**: We use BM25 (Robertson et al., 1995) to retrieve similar cases and their corresponding charges to provide context for the LLMs when making predictions. **Few-shot-SBERT**: We use Sentence-BERT (SBERT) (Reimers, 2019) to retrieve similar cases and their corresponding charges. The retrieval corpus is conducted using LJPIV-CAIL-Train dataset. Additionally, we directly fine-tuned the LLMs using the LJPIV-CAIL-Train, which we refer to as **Fine-Tuning-Direct**. For the implementation of trichotomous reasoning, we denote the methods as **Zero-shot-Tri** and **Fine-Tuning-Tri**, respectively. The link and license for the datasets and LLMs can be found in Appendix B.

5.1.3 Implementation Details

Our implementation utilizes Huggingface Transformers (Wolf et al., 2020) in the PyTorch framework. Considering the long length of legal documents and the consumption of computational resources, we chose to retrieve one example for few-shot retrieval. For Fine-Tuning-Tri, we used LoRA (Hu et al., 2021) to efficiently fine-tune the LLMs. We used the Adam optimizer (Kingma and Ba, 2014), set the initial learning rate to 5e-5, batch size to 16, used a cosine learning rate

[7]https://github.com/CSHaitao/LexiLaw

Dataset		LJPIV-CAIL (in-domain)				LJPIV-ELAM (cross-domain)				LJPIV-LeCaRD (cross-domain)			
Category	Model	Acc	P	R	F_1	Acc	P	R	F_1	Acc	P	R	F_1
Legal LLM	Disc-LawLLM	30.54	30.52	35.12	29.50	24.00	15.39	18.00	14.53	20.00	20.15	21.34	16.74
	LexiLaw	18.21	20.85	18.51	18.02	15.20	13.71	13.73	11.69	13.75	10.45	16.67	10.43
	fuzi.mingcha	19.46	25.16	20.83	20.72	13.40	13.07	13.54	11.16	10.00	9.23	9.96	7.59
Qwen2 (Qwen2-7B-Instruct)	Zero-shot	29.82	29.25	30.27	27.14	22.40	18.61	17.49	15.88	22.50	21.23	27.78	21.58
	Zero-shot-CoT	35.18	31.51	31.76	29.12	27.20	19.09	16.50	15.44	26.25	23.22	26.76	21.08
	Zero-shot-Tri (Ours)	50.71	35.90	34.75	32.68	38.60	22.18	19.80	18.46	42.50	**32.30**	**35.69**	**31.68**
	Few-shot-BM25	36.79	34.65	35.48	32.36	25.00	18.63	19.18	15.83	26.25	25.51	28.51	23.22
	Few-shot-SBERT	36.07	35.86	36.17	32.94	24.00	19.31	18.98	16.61	23.75	24.15	26.99	22.02
	Fine-Tuing-Direct	82.68	66.11	60.73	60.62	58.20	23.23	19.68	18.40	52.50	8.30	10.26	8.15
	Fine-Tuing-Tri (Ours)	**86.96**	**69.41**	**68.83**	**67.42**	**68.00**	**27.02**	**25.13**	**23.53**	**56.25**	23.90	19.73	20.22
Baichuan2 (Baichuan2-7B-Chat)	Zero-shot	24.46	25.54	24.11	21.93	20.60	15.12	14.53	12.85	16.25	12.62	16.45	11.57
	Zero-shot-CoT	27.68	29.28	22.33	23.48	23.00	17.54	15.69	14.55	17.50	12.00	11.76	9,27
	Zero-shot-Tri (Ours)	46.25	32.72	31.34	29.04	23.20	18.18	16.79	15.00	22.50	17.81	16.41	13.53
	Few-shot-BM25	27.68	29.28	22.33	23.48	23.20	16.29	15.35	13.46	25.00	13.41	21.45	13.77
	Few-shot-SBERT	35.18	29.69	27.21	26.00	25.00	17.11	14.97	13.20	23.75	14.45	21.89	14.92
	Fine-Tuing-Direct	82.32	67.41	60.77	61.37	62.00	25.10	20.86	19.48	52.50	7.35	8.84	7.89
	Fine-Tuing-Tri (Ours)	**87.32**	**70.76**	**68.08**	**66.76**	**65.00**	**28.22**	**23.58**	**22.32**	**63.75**	**27.95**	**23.19**	**23.63**

Table 3: Performance comparisons between Tri (Trichotomous) and the baselines on LJPIV-CAIL, LJPIV-ELAM and LJPIV-LeCaRD datasets. The best performance is indicated in bold, and the second best is underlined.

schedule, and fine-tuned for three epochs. All experiments are conducted on Nvidia A6000 GPUs. The code and datasets can be found at https://anonymous.4open.science/r/NG-467D.

5.2 Main Results

We conducted legal judgment prediction experiments on three datasets, and the results are shown in Table 3. From the table, we can draw the following conclusions:

• **Effectiveness of the Trichotomous Reasoning**. We found that across the three datasets and two open-domain base LLMs, the reasoning-based methods, Zero-shot-Tri and Fine-Tuning-Tri, consistently achieved encouraging results. This demonstrates the effectiveness of trichotomous reasoning in predicting convictions, as it enables the LLMs to consider both guilt and innocence, leading to more accurate predictions of innocence.

• **The Legal LLMs Perform Poorly**. It can be found that although the legal LLMs have been fine-tuned with a substantial amount of legal knowledge, the performance of the three legal LLMs in predicting verdicts is relatively poor, even inferior to the zero-shot performance of open-domain LLMs. This is because the legal LLMs were fine-tuned on datasets with guilty legal judgments only and have not been exposed to cases of innocence. Therefore, when given a case, they do not recognize the possibility of rendering a not-guilty verdict.

• **Cross Domain Results**. We can observe that when transferring the LLM fine-tuned on the LJPIV-CAIL dataset to the other two datasets, the improvement is not as significant as the improvement on the CAIL dataset itself. This is due to the different sources of cases in different datasets, which may result in variations in length and style (for instance, as shown in Table 2, the case length in ELAM is twice that of CAIL). The differences between datasets lead to variations in the effectiveness of judgment prediction.

Category	Model	Acc	P	R	F_1
Qwen2	Tri	**86.96**	**69.41**	**68.83**	**67.42**
	w/o Level 3	77.86	59.83	61.01	58.70
	w/o Level 2	69.29	57.53	59.21	56.64
Baichuan2	Tri	**87.32**	**70.76**	**68.08**	**66.76**
	w/o Level 3	78.21	62.83	61.22	59.04
	w/o Level 2	68.75	59.91	60.42	57.44

Table 4: Ablation studies for trichotomous reasoning on LJPIV-CAIL. Tri refers to Fine-Tuing-Tri.

5.3 Ablation Study on Trichotomous Levels

To explore the effectiveness of each level in the trichotomous reasoning, we investigated the performance of two LLMs on the LJPIV-CAIL test set by progressively removing the third and second levels from the Fine-Tuning-Tri. The results are shown in Table 4, and we provide a detailed analysis below:

• **w/o Level 3**. This indicates the removal of the level of culpability from the entire reasoning process, which means not considering the three situations mentioned in Sec. 3.2 in rendering a not-guilty verdict. The performance drop observed in both LLMs highlights the importance of the third level.

• **w/o Level 2**. This indicates the removal of the level of Unlawfulness on top of removing level 3, meaning not considering not-guilty verdicts due to the defendant's actions being a result of self-defense or necessity. The decrease in prediction results observed in both LLMs underlines the importance of the second level. Allowing LLMs to judge whether the defendant's actions were justified by self-defense or necessity can also improve

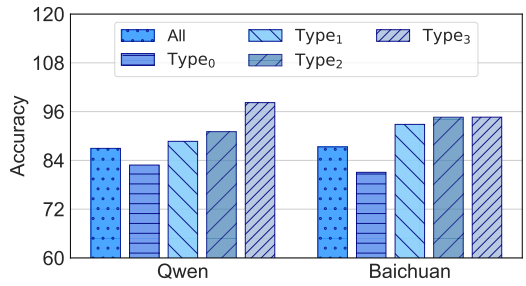


Figure 4: Prediction accuracy for different case types on the LJPIV-CAIL test set. "All" indicates the overall accuracy, while "Type$_0$" represents the accuracy for guilty cases. "Type$_1$", "Type$_2$", and "Type$_3$" represent the accuracies for non-guilty cases due to lack of elements, unlawfulness, and culpability, respectively.

the accuracy of not-guilty verdicts.

These findings suggest that each level in the trichotomous reasoning has a significant role in improving the accuracy of legal judgment predictions.

5.4 Guilty vs. Non-Guilty Predictions

The dataset contains both guilty and not guilty cases, with three types of not guilty verdicts: due to elements of offense, unlawfulness, and culpability. To evaluate how accurately the LLM predicts guilty and not guilty outcomes, we analyzed the final verdicts of two LLMs on the LJPIV-CAIL test set, categorizing them into four types: guilty (Type$_0$), not guilty due to elements of offense (Type$_1$), not guilty due to unlawfulness (Type$_2$), and not guilty due to culpability (Type$_3$). We then compared their predictive accuracies with the overall accuracy (All).

As shown in Figure 4, both LLMs exhibited similar trends. The prediction accuracy for not guilty cases (Type$_1$, Type$_2$, Type$_3$) was higher than the overall accuracy (All), while the accuracy for guilty cases (Type$_0$) was lower. This suggests that the trichotomous reasoning is particularly effective for predicting not-guilty verdicts. When comparing the three types of not-guilty verdicts, we observed that the accuracy for cases not guilty due to elements was lower than those for unlawfulness and culpability. This difference arises because these characteristics vary in complexity. Elements require LLMs to differentiate between multiple reasons for not guilty verdicts across various charges, whereas unlawfulness and culpability are more generalizable.

Additionally, we analyzed the performance of the best-performing legal LLM, DISC-LawLLM, on the CAIL test set. We found that its accuracy for guilty verdicts was 61.07%, while the accuracy for all three types of not-guilty verdicts was 0%.

Table 5: The performance of larger models on LJPIV. Tri is based on Qwen-7B.

Model	Acc	P	R	F$_1$
GPT-4o	26.61	24.28	24.44	22.21
Deepseek-R1	32.86	39.07	41.71	37.76
Deepseek-R1 (CoT)	33.93	45.57	43.43	42.62
Zero-shot-Tri	50.71	35.90	34.75	32.68
Fine-Tuning-Tri	**86.96**	**69.41**	**68.83**	**67.42**

This is likely due to overfitting during fine-tuning, causing the model to favor guilty verdicts for given cases, as shown in Figure 2.

5.5 The Performance of larger models

In this section, we evaluate the performance of GPT-4o and DeepSeek-R1 (Guo et al., 2025) on the LJPIV dataset. We instructed GPT-4o and DeepSeek-R1 to determine the suspect's charges based on the Chinese Criminal Law and the case facts. "CoT" refers to we prompted DeepSeek-R1 to specifically consider the possibility of a non-guilty verdict. The results are shown in Table 5. We can see that Fine-Tuning-Tri still achieved optimal performance.

Upon analyzing the cases where DeepSeek performed poorly, we found that it struggles with accurately understanding legal definitions, such as "necessary limits" and "serious harm"—for example, whether minor injuries qualify as "serious harm". Additionally, we observed that DeepSeek tends to over-rely on general legal practices of "strict recognition" while overlooking the specific details of individual cases.

6 Conclusion

In this paper, we introduce LJPIV, the first benchmark dataset for legal judgment prediction with innocent verdicts. We extend three widely-used legal datasets through LLM-based augmentation and manual verification. We further introduce zero-shot prompt-based and fine-tuning methods to equip open-domain LLMs with trichotomous reasoning capabilities, particularly for predicting innocent outcomes in legal judgment prediction. Extensive experiments reveal that (1) current legal LLMs have significant room for improvement, with even the best models achieving an F1 score below 0.3 on LJPIV; (2) fine-tuning on LJPIV substantially improves both in-domain and cross-domain judgment prediction accuracy for open-domain LLMs, especially in cases resulting in an innocent verdict; and (3) our trichotomous reasoning strategies further enhance the legal judgment performance.

7 Limitations

Legal systems across the world vary significantly, and different systems often adhere to distinct doctrines when making legal judgments. For instance, common law systems rely heavily on case precedents, whereas civil law systems are based on codified statutes. As a result, this paper focuses solely on Chinese datasets within the civil law system. Future work will aim to adapt our approach to other legal systems, including common law jurisdictions, and explore datasets in other languages to increase the generalizability and applicability of our approach across different legal contexts.

8 Ethical Considerations

We demonstrate the legality and compliance of our data construction process from the aspects of data anonymization, licensing and usage, legal compliance, and human oversight:

Data Anonymization: We use publicly available legal LJP datasets (Ma et al., 2021; Yu et al., 2022b; Xiao et al., 2018), which explicitly state that all personal information has been anonymized. According to China's Personal Information Protection Law, anonymized data is no longer legally protected. Therefore, our dataset fully complies with these regulations.

Licensing and Usage: As outlined in the Appendix B, the datasets we use are protected by the MIT License, which explicitly permits free use, reproduction, and modification. Our usage is limited to research purposes. Based on these points, we believe our data construction method is both legal and compliant.

Legal Compliance: The modifications generated by the LLM for the dataset are executed in strict adherence to legal principles and standards. The integrated facts follow normative legal reasoning.

Human Oversight: All modifications to the dataset undergo manual review to ensure consistency with regulatory and legal requirements.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 22–31.

Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. 2021. Chatgpt goes to law school. *J. Legal Educ.*, 71:387.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Chenlong Deng, Zhicheng Dou, Yujia Zhou, Peitian Zhang, and Kelong Mao. 2024. An element is worth a thousand words: Enhancing legal case retrieval by incorporating legal elements. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2354–2365.

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.

Markus Dirk Dubber. 2005. Theories of crime and punishment in german criminal law. *The American journal of comparative law*, 53(3):679–707.

Roni A Elias. 2015. Three cheers for three tiers: Why the three-tier system maintains its legal validity and social benefits after granholm. *DePaul Bus. & Comm. LJ*, 14:209.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

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

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction: A survey of the state of the art. In *IJCAI*, pages 5461–5469.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured nlp tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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. *arXiv preprint arXiv:2106.09685*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023. On the (in) effectiveness of large language models for chinese text correction. *arXiv preprint arXiv:2307.09007*.

Jinliang Lu, Chen Wang, and Jiajun Zhang. 2024. Diver: Large language model decoding with span-level mutual information verification. *arXiv preprint arXiv:2406.02120*.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.

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

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*.

Weicong Qin, Zelin Cao, Weijie Yu, Zihua Si, Sirui Chen, and Jun Xu. 2024. Explicitly integrating judgment prediction with legal document retrieval: A law-guided generative approach. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2210–2220.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. *arXiv preprint arXiv:2407.21065*.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348.

ZhongXiang Sun, Kepu Zhang, Weijie Yu, Haoyu Wang, and Jun Xu. 2024. Logic rules as explanations for legal case retrieval. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10747–10759.

Pengfei Wang, Ze Yang, Shuzi Niu, Yongfeng Zhang, Lei Zhang, and ShaoZhang Niu. 2018. Modeling dynamic pairwise attention for crime classification over legal articles. In *the 41st international ACM SIGIR conference on research & development in information retrieval*, pages 485–494.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Shiguang Wu, Zhongkun Liu, Zhen Zhang, Zheng Chen, Wentao Deng, Wenhao Zhang, Jiyuan Yang, Zhitao Yao, Yougang Lyu, Xin Xin, Shen Gao, Pengjie Ren, Zhaochun Ren, and Zhumin Chen. 2023. fuzi.mingcha. https://github.com/irlab-sdu/fuzi.mingcha.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

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*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022a. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

Weijie Yu, Zhongxiang Sun, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen. 2022b. Explainable legal case matching via inverse optimal transport-based rationale extraction. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 657–668.

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

Yating Zhang, Yexiang Wang, Fei Cheng, Sadao Kurohashi, et al. 2023. Reformulating domain adaptation of large language models as adapt-retrieve-revise. *arXiv preprint arXiv:2310.03328*.

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.

Youchao Zhou, Heyan Huang, and Zhijing Wu. 2023. Boosting legal case retrieval by query content selection with large language models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 176–184.

A More details of the annotators

We employed a total of five legal annotators, all of whom have legal education backgrounds and are familiar with the cases in the dataset. The team consisted of:

- One judicial expert: A female legal scholar aged between 40–50, holding a Ph.D. in law with extensive expertise in criminal law and judicial practice.
- Four postgraduate students specializing in criminal law: The group included two males and two females aged between 23–30, each with over five years of experience studying criminal law theory and practical exposure to judicial procedures.

The annotation process was carefully guided by the judicial expert, who established the annotation guidelines and ensured consistency across the dataset. We informed the annotators that their annotated data would be used for scientific research and provided them with fair compensation based on local standards. During each iteration, the expert resolved any disagreements or ambiguous cases encountered by the annotators, making the final decisions to ensure the annotations were both accurate and aligned with the intended legal framework.

B More Details of Datasets and Models

In this section, we provide the link and license for the dataset we used, as shown in Table 6.

Type	Dataset	URL	Licence
Dataset	CAIL-2018	https://github.com/china-ai-law-challenge/CAIL2018	MIT License
	LeCaRD	https://github.com/myx666/LeCaRD	MIT License
	ELAM	https://github.com/ruc-wjyu/IOT-Match	MIT License
LLM	LexiLaw	https://github.com/CSHaitao/LexiLaw	MIT license
	DISC-LawLLM	https://github.com/FudanDISC/DISC-LawLLM	Apache-2.0 license
	fuzi.mingcha	https://github.com/irlab-sdu/fuzi.mingcha	Apache-2.0 license
	Qwen2-7B-Instruct	https://huggingface.co/Qwen/Qwen2-7B-Instruct	Apache-2.0 license
	Baichuan2-7B-Chat	https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat	Apache-2.0 license

Table 6: The URLs and licenses for the datasets and LLMs used by LJPIV.