CLEAR: A Framework Enabling Large Language Models to Discern Confusing Legal Paragraphs

Qi Xu[1], **Qian Liu**[2], **Hang Yu**[1], **Hao Fei**[3] **Shuhao Guan**[4], **Xiao Wei**[1*]

[1]School of Computer Engineering and Science, Shanghai University, China
[2]School of Computer Science, University of Auckland, New Zealand
[3]School of Computing, National University of Singapore, Singapore
[4]School of Computer Science, University College Dublin, Ireland
{welch,yuhang,xwei}@shu.edu.cn, Liu.Qian@auckland.ac.nz

Abstract

Most of the existing work focuses on enabling LLMs to leverage legal rules (*e.g.*, law articles) to tackle complex legal reasoning tasks, but largely overlooks their ability to understand legal rules. To better evaluate the LLMs' capabilities in this regard, in this work, we propose a new challenge task: Legal Paragraph Prediction (LPP), which aims to predict the legal paragraph given criminal facts. Moreover, to enhance the legal reasoning ability of LLMs, we propose a novel framework CLEAR, enabling LLMs to analyze legal cases with the guidance of legal rule insights. CLEAR consists of four key components, where the *Legal Rules Retriever* aims to retrieve legal rule knowledge, and the *Rule Insights Generator* is used to generate legal insights guiding the LLM's reasoning, then the *Case Analyzer* analyze the case with the guidance of legal rule insights given criminal facts. Finally, the *Legal Reasoner* synthesizes the criminal facts, legal rule insights, and analysis results to derive the final decision. By conducting extensive experiments on a real-world dataset, experimental results validate the effectiveness of our proposed model. Our codes and dataset are available at https://github.com/xuqi220/CLEAR.

Warning!!! This paper contains discussions of violent criminal activities. Discretion is advised.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance across various Natural Language Processing (NLP) tasks (DeepSeek-AI et al., 2025; Qwen et al., 2025; OpenAI, 2023; Fei et al., 2024a,b, 2025; Wu et al., 2024a,b). However, prior studies show that they often struggle to capture the complexities of real-world applications, particularly in specialized domains such as law (Mishra et al., 2025; Jiang et al., 2024; Kang et al., 2023).

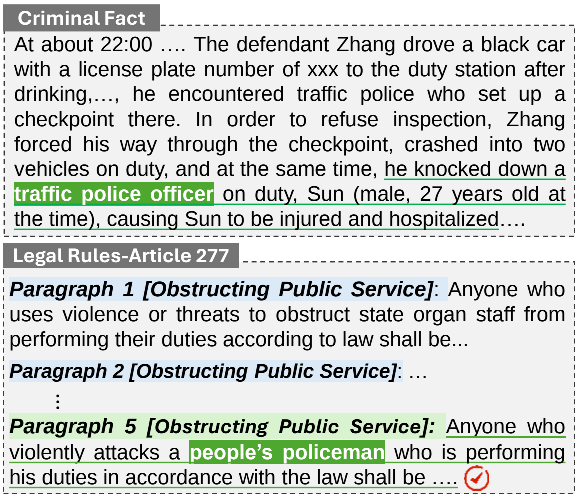

Figure 1: An example of the criminal fact and legal paragraphs (`Paragraph 1 of Article 277` and `Paragraph 5 of Article 277`) related to the crime of `Obstructing Public Service`. The subtle difference between the two lies in whether the criminal act targets the policeman. `Paragraph 5 of Article 277` is more aligned with the given criminal fact.

In this work, we further investigate the efficacy of LLMs in realistic legal contexts, where precision, contextual understanding, and domain expertise are critical (Yuan et al., 2024).

For legal professionals, a fundamental competency is accurately aligning legal rules with given criminal facts (MacCormick, 2005). Within a legal system, such as the Chinese Criminal Law System, legal rules refer to explicitly defined, codified legal provisions encompassing law articles and their paragraphs. A key challenge arises from the highly nuanced distinctions between certain legal paragraphs, which may be so subtle that even experienced legal practitioners struggle to differentiate them. For example, as shown in Figure 1, both the `Paragraphs 1 and 5` of `Article 277` manifest the criminal circumstance of the crime of `Obstructing Public Service`. The nuanced distinction lies in whether the criminal acts target a policeman. Therefore, properly applying legal paragraphs reflects LLMs' abilities to understand

* corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8937–8953
November 4-9, 2025

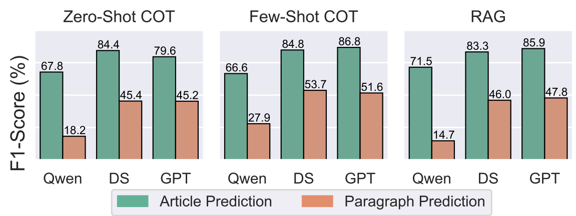

Figure 2: The performance of LLMs on article prediction and paragraph prediction under the chain of thought (COT) and retrieval-augmented generation (RAG) methods, respectively. The Qwen, DS, and GPT represent the Qwen2.5-14B, DeepSeek-v3, and GPT-o1, respectively.

legal rules and criminal facts. Recently, many researchers have enabled LLMs to understand and use legal rules to address complex legal reasoning tasks (Xu et al., 2024b; Jiang et al., 2024; Wu et al., 2024c; An et al., 2022). As such, a natural question arises: *Are large language models capable of reasoning applicable legal rules?*

To address this inquiry, we sample real-world criminal cases and ask LLMs to predict the applicable law article and paragraph. As shown in Figure 2, LLMs achieve promising results on the law article prediction task but struggle with paragraph prediction, which indicates the necessity for LLMs to possess not only complex reasoning abilities but also refined skills in detecting and discriminating between subtle legal elements. To better evaluate LLMs' capability of legal reasoning, we propose a challenging task: Legal Paragraph Prediction (LPP), a more nuanced legal rule prediction task detailed in Section 4. The task contributes to tackling a fundamental research question: *How to enable LLMs to differentiate legal paragraphs?*

Generally, LLMs consistently face substantial challenges when performing legal reasoning tasks, demonstrating notable limitations in the legal analysis scenario: (1) **Deficient Domain Knowledge**. Their dependence on parametric knowledge alone frequently leads to hallucinatory outputs (Dahl et al., 2024; Deroy et al., 2023). (2) **Distracted by Irrelevant Information**. The content of legal documents is often lengthy (Xiao et al., 2021), and LLMs can easily be misled by irrelevant information, leading to erroneous conclusions. (Shi et al., 2023). (3) **Ignoring Factual Details**. The disregard for nuanced case particulars, which encapsulate essential legal elements, compromises the ability to derive accurate legal conclusions (Xu et al., 2024b; Deng et al., 2024).

To tackle these issues, we introduce a novel framework <u>C</u>omplex <u>L</u>egal Cas<u>E</u> <u>A</u>nalyze<u>R</u> (CLEAR) which aims to improve LLMs' ability of distinguish confusing legal paragraphs. Specifically, for given criminal facts, a *Legal Rules Retriever* first identifies relevant legal paragraphs that may cause confusion for the LLM and retrieves their definitions. Secondly, they are assigned to *Rule Insights Generator* for nuanced interpretation. Thirdly, a *Case Analyzer* processes these interpretations along with the criminal facts to elicit LLM focus on the case details defined in the legal paragraphs to generate legal case insights. Finally, a *Legal Reasoner* synthesizes the criminal facts and legal insights to derive a conclusion.

To evaluate the effectiveness of our framework CLEAR, we conduct extensive experiments on the LPP dataset. The results show that CLEAR achieves 8.26% and 5.83% improvements in terms of precision and recall compared with previous advanced methods. To evaluate the quality of legal insights produced by our CLEAR, we implement the Legal Reasoner module by fine-tuning LMs, which further improves the performance of our framework.

Our contributions can be summarized as follows:

- We introduce a new benchmark LPP, focusing on legal paragraph prediction, to better evaluate LLMs' capabilities of legal reasoning.
- We present a novel framework CLEAR designed to distinguish confusing legal paragraphs, which enables LLMs to focus on legal case details to improve the performance.
- The experiment results show that our method surpasses the previous state-of-the-art (SOTA) methods by a large margin. Further study shows the high quality of the legal insights generated by our framework, indicating the effectiveness of our CLEAR. Our contributions thus offer a robust foundation for advancing AI's role in the legal domain.

2 Related Works

LLMs in Legal AI. In recent years, LLMs, such as DeepSeek (DeepSeek-AI et al., 2025), Qwen (Qwen et al., 2025), and GPT-4 (OpenAI, 2023) have shown remarkable performance in various complex tasks such as logical reasoning (Xu et al., 2024a; Fei et al.) and mathematical reasoning (Shao et al., 2024). In the law domain, many researchers explore how to design efficient prompts for LLMs to complete legal reasoning tasks (Trautmann et al., 2022; Yu et al., 2023; Shui et al., 2023;

Barron et al., 2025; Kant et al., 2025). For example, Yu et al. (2023) and Wu et al. (2023) enable LLMs to conduct legal reasoning like a legal expert. Some works focus on specializing LLMs in the legal domain (Huang et al., 2023; Li, 2023; Liu et al., 2023). For example, He et al. (2023) fully pre-trained an LLM with 7 billion parameters on the legal dataset. Yue et al. (2024) fine-tune LLMs for the court view generation task. Recently, enhancing LLMs with legal rules has attracted substantial attention (Wu et al., 2023; Yuan et al., 2024; Xu et al., 2024b; Luo et al., 2025). Despite advances, most of these methods depend on integration of law articles into LLMs, disregarding their intrinsic ability for legal paragraph reasoning. To fill this gap, in this study, we focus on evaluating and improving the ability of LLM in the legal paragraph reasoning task.

Legal Rule Understanding. Legal rules, being fundamental to legal systems, also hold significant importance in AI-driven legal tasks such as Legal Judgment Prediction (Xiao et al., 2018; Chalkidis et al., 2019), Legal Question Answering (Zhong et al., 2020a; Yao et al., 2025), and Court View Generation (Yue et al., 2024). Recently, many works focus on the law article prediction task (Zhong et al., 2019; Chalkidis et al., 2020; Fei et al., 2024c; Hwang et al., 2022; Luo et al., 2023, 2025) or taking law articles as knowledge to improve the performance of LLM (An et al., 2022; Wu et al., 2023; Cui et al., 2023; Yuan et al., 2024). Despite significant progress, these studies neglect the crucial dimension of the more nuanced legal paragraph interpretation, particularly the differentiation of ambiguous legal provisions, rendering current advanced methodologies inadequate. To fill this gap, we propose a new challenge task: Legal Paragraph Prediction (LPP). Unlike the law article prediction tasks (Xiao et al., 2018; Lyu et al., 2023), our work is principally devoted to thoroughly understanding legal rules, specifically addressing the challenge of distinguishing confusing legal paragraphs.

3 LPP Dataset

Data Collection. Due to the absence of the paragraph prediction dataset, to fill this gap, we construct our dataset LPP. Cases from LPP are sourced from the government website China Judgment Online (CJO) [1], which is widely used in the LegalAI tasks (Xiao et al., 2018; Xu et al., 2024b). Following previous work (Xiao et al., 2018; Wei et al., 2024), we first extract the defendants' names, criminal facts, and legal rules (law articles and paragraphs) from the case documents automatically. Then, we keep the cases that contain only one law article. We exclude cases with criminal facts containing fewer than 150 characters. Paragraphs with a frequency of less than 30 are removed. Finally, we obtain 20,040 cases covering 20 law articles and 47 paragraphs. In addition, we also construct a legal knowledge base, detailed in the Appendix A, derived from authoritative law books (Zhong et al., 2020a). It is denoted as $\mathcal{K} = \{k_{i1}, k_{i2}, ..., k_{im}\}$, where k_{ij} represents the definition of the j-th paragraph of the i-th law article.

Items	Train Set	Dev. Set	Test Set	Total
Avg. Length of Fact	2,057.8	2,046.0	2,066.1	2,057.1
# Cases	11,981	4,001	4,022	20,004
# Paragraphs	47	47	47	47

Table 1: Statistics of our dataset, where the "#", "Dev." and "Avg." denote "the number of", "Development" and "Average", respectively.

Quality Management. To ensure the quality of LPP, we employ 20 senior Ph.D. students of Law. Each of them is equipped with extensive legal expertise and a thorough comprehension of the Criminal Law of China. Specifically, given the criminal facts and the automatically annotated results, we ask the annotators to verify and rectify any inaccuracies in the automated annotations. Annotators are instructed to dedicate a minimum of 10 minutes per factual case and receive compensation at a rate of $25 per hour, calculated according to the actual time spent completing the annotations. Moreover, we first split the dataset into 20 segments. Annotators are assigned 3 randomly selected segments, and each segment undergoes triple-independent review. Secondly, cases with inconsistent annotations undergo deliberative discussion among annotators, followed by re-annotation until unanimous agreement is achieved. Finally, we exclude cases where consensus cannot be achieved during the second step, thereby ensuring only consistently annotated data remains.

Dataset Statistics. After the annotation process, as shown in Table 1, there are 20,004 cases covering 20 law articles and 47 paragraphs. We split the dataset into a train set, development set, and test set following the ratio of 3:1:1. Please refer to Appendix A for more details.

[1] https://wenshu.court.gov.cn/

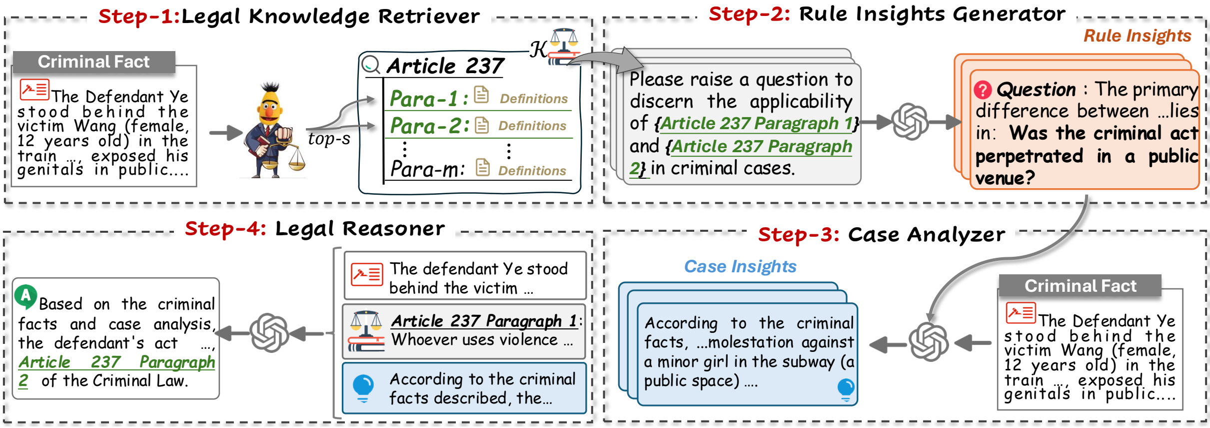

Figure 3: Overall architecture of CLEAR, which consists of four components: the *Legal Knowledge Retriever*, the *Rule Insights Generator*, the *Case Analyzer*, and the *Legal Reasoner*. Specifically, given the criminal fact, the *Legal Knowledge Retriever* first to provides candidate legal rules and their definitions. Then the *Rule Insights Generator* followed by the *Case Analyzer* is applied to generate insights of cases and rules. Finally, the *Legal Reasoner* aggregates criminal facts, legal rules, and case insights to derive the conclusion.

4 Methodology

4.1 Overview

Task Definition. We first introduce the definition of the Legal Paragraph Prediction (LPP) task. **Fact** denotes the description of a criminal case involving the defendant's background information (*e.g.*, previous convictions) and the criminal facts, which are denoted as $f = \{w_1, w_2, ..., w_{|f|}\}$ where w_i is the i-th word. **Legal Rule** refers to a collection of law articles and paragraphs, where each law article contains one or more paragraphs. We denote the law article set as $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ and denote paragraphs under the a_i as $\mathcal{P}_i = \{p_{i1}, p_{i2}, ..., p_{im}\}$, where the a_i denote the i-th law article and the p_{ij} denote the j-th paragraph belongs to the i-th law article. n and m denote the number of the law articles and paragraphs, respectively. We take **Legal Rule Definitions** as the legal knowledge denoted as $\mathcal{K} = \{k_{ij} | i \in [1, n]; j \in [1, m]\}$, where the k_{ij} represents formal definitions for paragraph p_{ij} detailed in Appendix A. This work, given the fact f, aims to automatically predict the paragraph p_{ij}.

4.2 Framework CLEAR

The overall architecture of our CLEAR framework is shown in Figure 3, which comprises four modules: (1) The *Legal Knowledge Retriever* predicts candidate paragraphs and extracts their corresponding definitions. (2) Then, the *Rule Insights Generator* formulates distinguishing questions. (3) The *Case Analyzer* answers these questions based on criminal facts to produce case insights. (4) Finally, the *Legal Reasoner* integrates the criminal facts, legal rules, and case insights to generate the final decision. The basic idea centers on utilizing insights gained from meticulous analysis of nuanced legal rules' definitions to forecast the application of the specific legal rule.

Legal Knowledge Retriever. As aforementioned, LLMs may generate hallucinations due to reliance only on parametric knowledge (Dahl et al., 2024). To alleviate this problem, we design a Legal Knowledge Retriever module to predict candidate legal paragraphs and retrieve their definitions. Specifically, given a criminal case fact, the module first obtains the representation of the whole fact description f by passing it into the pre-trained encoder (*e.g.*, LegalBERT (Zhong et al., 2019)):

$$\mathbf{h}_f = \mathrm{Enc}(\texttt{[CLS]}\ f\ \texttt{[SEP]}), \tag{1}$$

where [CLS] and [SEP] are the special tokens. The $\mathbf{h}_f$ is the output embedding of the token [CLS]. Then, the linear classifier and softmax function are applied to predict the probabilities of the paragraphs $\hat{\mathbf{p}} \in \mathbb{R}^{N_p}$, where N_p is the number of paragraphs in our LPP. We train the module by optimizing the paragraph classification loss:

$$\mathcal{L} = \mathbb{E}[-\sum_{n_p=1}^{N_p} \mathbf{y}(n_p|\mathbf{h}_f) \log(\hat{\mathbf{p}}(n_p|\mathbf{h}_f))], \tag{2}$$

where $\mathbb{E}$ represents the average expectation. The $\mathbf{y}$ denotes the ground-truth probability. If n_p is the ground-truth, $\mathbf{y}$ equal to 1; otherwise 0. During the inference stage, we select the *top*-s paragraphs as candidate paragraphs. Then, we retrieve their definitions as legal knowledge.

Rule Insights Generator. After obtaining the candidate paragraphs $\{p_{1\sim s} \in \mathcal{P}\}$ and their def-

Prompt for Rule Insights Generator

#Definitions:# The definition of { p_i } and { p_j } is described as following: { p_i } : { k_i } ; { p_j } : { k_j }
#Instruction:# Please raise questions to discern the applicability of { p_i } and { p_j }, in the real judicial scenario.

Figure 4: Prompting template for generating rule insights, where p and k represent the paragraph and its definition, respectively.

initions $\{k_{1\sim s} \in \mathcal{K}\}$, the module aims to generate questions to reveal their difference. Specifically, we ask the LLM to raise questions to disambiguate each pair of paragraphs $< p_i, p_j >$ from $p_{1\sim k}$ when confronting the real judicial scenario. The prompting template is shown in Figure 4. The prompt begins with the paragraph pair $< p_i, p_j >$ and their definitions which are followed by the instructions to elicit the LLM to generate a question $q_{<p_i,p_j>}$ for distinguishing the nuanced legal rules. By producing insights that compare nuanced legal paragraphs, this module directs the model's analytical focus and fosters a deeper analysis of how legal rules function in real-world cases, enabling the distinction of easily confusable legal paragraphs.

Prompt for Case Analyzer

#Criminal Facts:# The criminal facts is { f } :
#Question:# {$q_{<p_i,p_j>}$}
#Instruction:# Please answer question based on the Criminal Facts.

Figure 5: Prompting template for generating case insights. where f and $q_{<p_i,p_j>}$ represent the criminal facts and question generated by the rule insights generator module, respectively.

Case Analyzer. As mentioned in Section 1, LLMs can be easily distracted by irrelevant context (Shi et al., 2023). To address this issue, this module asks LLMs to analyze the case with the guidance of rule insights generated by the rule insights generator. Specifically, for each question $q_{<p_i,p_j>}$, we ask the LLM to generate the answer $Ans_{<p_i,p_j>}$ based on the given criminal facts. The prompt template is shown in Figure 5. The prompt begins with the criminal facts and a question, followed by the instruction to elicit the LLM to generate the answer. Under question-driven prompting, the LLM prioritizes legally salient details (*e.g.*the victim's identity in Figure 1) that disambiguate overlapping legal elements, refining its outputs and boosting confidence in its reasoning.

Prompt for Legal Reasoner

#Criminal Facts:# The criminal facts is { f } :
#Case Insights:# {$Ans_{<p_i,p_j>}$}
#Legal Rules:# { p_i } : { k_i } ; { p_j } : { k_j }
#Instruction:# Please select the applicable legal rule based no the criminal facts and case insights.

Figure 6: Prompting template for the Legal Reasoner. where f, $Ans_{<p_i,p_j>}$, p, and k represent the criminal facts, insights, paragraph, and definition, respectively.

Legal Reasoner. As a consequence of neglecting contextual case details in the juridical analysis, the LLM may derive unsatisfactory legal conclusions (Deng et al., 2024). To tackle the issue, for each case, we integrate the legal paragraphs, their definitions, and legal insights to enable LLMs to conduct legal reasoning focusing on the key case details. Specifically, we implement this module using two strategies. **(1) Prompt-Based LLMs**, which makes the final decision by prompting LLMs. Specifically, we concatenate the legal knowledge and legal insights to ask the LLM to generate an applicable paragraph. Figure 6 shows the prompt template. It begins with the criminal fact, case insights, and candidate legal rules linked with their definitions, followed by the instructions to ask the LLM to make a final decision. (2) **SFT-Based LM**, which implements the legal reasoner module by fine-tuning a language model such as LegalBERT (Zhong et al., 2019). To construct a dataset for legal reasoner development, we execute a pipeline that combines the above-mentioned modules and the prompt-based legal reasoner based on Qwen2.5-14B (Qwen et al., 2025) to the training and development set, extracting legal rules, case insights, and final conclusions. Then, we eliminate the sample that contains incorrect conclusions and obtain the fine-tuned dataset. Finally, we train the legal reasoner module on the distilled data following Section 4.2 the procedure regarding the training paragraph prediction.

5 Experiment

5.1 LLMs

We first introduce LLMs used in our study. **GPT-o1** (OpenAI, 2023), a powerful reasoning model for solving complex reasoning tasks. The version of `o1-preview` is used, and it is available from the OpenAI API [2]. **DeepSeek** (DeepSeek-AI et al., 2025), a strong Mixture-of-Experts (MoE) language model with 671B parameters. The version of `deepseek-v3` is used, and it is available from the

[2]https://openai.com/

deepseek platform [3]. **Qwen2.5-14B** (Qwen et al., 2025) an open-source [4] language model with 14B parameters. **Lexi-Law** (Li, 2023), **DISC-LLM** (Yue et al., 2023), which are specialized LLMs in the law domain based on GLM2-6B (Du et al., 2022) and BaiChuan2-13B (BaiChuan-Inc, 2023).

5.2 Dataset and Evaluation Metrics

We evaluate our method on our dataset LPP, which is detailed in Section 3. Due to the large size of the dataset (>4k for the test set of LPP), following Shui et al. (2023); Wei et al. (2024), we randomly select a small balanced set from the test set of LPP, which contains 667 cases. We denote the small set sampled as SLLP. To evaluate the performance of our CLEAR and baseline models, we adopt the accuracy (Acc.), macro precision (Prec.), macro recall (Rec.), and macro F1 score as metrics.

5.3 Experiment Setting

Baseline Models. We compare our CLEAR with the following four groups of baseline models: **(1) Task-Specific Models** that are designed for the legal rule classification task: LADAN (Xu et al., 2020) and LEMM (Zhong et al., 2020b), which distinguish confusing law articles by using their definitions; EMP (Feng et al., 2022), which predicts legal rules by integrating legal events; NeurJudge (Yue et al., 2021), which predicts legal rules by splitting fact description separations. **(2) SFT-Based Language Models** that are implemented by fine-tuning language models: RoBerta (Liu et al., 2019), a robust bert-based language model; LegalBERT (Zhong et al., 2019) and Lawformer(Xiao et al., 2021), pretrained bert-based models on a large-scale legal dataset; Qwen2.5-14B (Qwen et al., 2025), which is fine-tuned by using LoRA (Hu et al., 2022). **(3) Prompt-Based LLMs** that are implemented based on advanced prompt-based methods: zero-shot COT (Kojima et al., 2023) and few-shot COT (Wei et al., 2023), which elicit LLMs to break down complex problem by think step-by-step. Retrieval-Augmented Generation (RAG) (Shui et al., 2023), which enrich the prompt by using the retrieved similar cases. **(4) Hybrid Approach**, which combines LLM and SLM capabilities (Wu et al., 2023) through two distinct strategies: **Predicted**, having the LLM select answers exclusively from the LegalBERT's top-4 predicted candidate paragraphs, and **Hierarchy**, limiting choices to paragraphs within the predicted legal articles.

Setting	BackBone Model	Metrics (%)			
		Acc.	Prec.	Rec.	F1
Task Specific	LADAN	41.17	39.40	38.79	38.40
	LEMM	48.63	41.66	41.10	41.23
	EPM	40.29	38.85	37.65	37.64
	NeurJudge	44.96	43.88	43.38	43.48
SFT	LegalBERT	64.02	60.78	59.52	53.98
	LawFormer	<u>64.52</u>	59.90	58.35	52.71
	RoBERTa	6147	52.17	57.08	49.56
	Qwen2.5-14B	41.33	33.34	40.96	36.76
Zero-Shot COT	GPT-o1	49.69	47.02	46.40	45.25
	DeepSeek-V3	51.42	46.12	48.66	45.37
	Qwen2.5-14B	22.94	16.60	21.72	18.21
	Lexi-Law	21.80	19.85	19.09	19.25
	DISC-Law	15.50	14.44	13.42	14.25
Few-Shot COT	GPT-o1	51.10	53.77	55.22	51.61
	DeepSeek-V3	52.37	53.48	54.63	53.66
	Qwen2.5-14B	28.68	27.09	29.82	27.90
	Lexi-Law	21.63	18.99	20.33	19.06
	DISC-Law	20.37	17.38	19.52	17.89
RAG	GPT-o1	48.40	45.87	49.57	47.78
	DeepSeek-V3	48.26	46.15	47.31	46.01
	Qwen2.5-14B	17.84	14.32	16.78	14.71
	Lexi-Law	9.60	9.88	10.46	10.92
	DISC-Law	10.18	9.97	9.91	9.85
Hybrid (Predicted)	GPT-o1	63.63	<u>61.65</u>	62.18	61.77
	DeepSeek-V3	64.07	60.68	62.72	61.45
	Qwen2.5-14B	44.08	41.41	42.11	40.16
Hybrid (Hierarchy)	GPT-o1	61.68	60.18	60.44	60.26
	DeepSeek-V3	63.52	61.02	<u>62.85</u>	<u>62.59</u>
	Qwen2.5-14B	44.23	41.11	42.52	39.76
CLEAR-PT (Impr.)	GPT-o1	69.72 (+5.20)	**69.49** (+7.84)	68.53 (+5.68)	67.21 (+4.62)
	DeepSeek-V3	**70.01** (+5.49)	68.99 (+7.34)	**68.68** (+5.83)	**67.30** (+4.71)
	Qwen2.5-14B	66.87 (+2.35)	65.39 (+3.74)	64.75 (+1.90)	63.39 (+0.8)
CLEAR-SFT (Impr.)	Qwen2.5-14B + LegalBERT	68.37 (+3.85)	69.91 (+8.26)	66.15 (+3.30)	65.07 (+2.48)

Table 2: The overall performance on the LPP. The best scores of previous methods are <u>underlined</u> and the SOTA performances are marked with **bold**. The "Impr." and "PT" represent the improvements and Prompt, respectively. We mark the improvements over 5% in red.

Implement Details. For our CLEAR, the Legal Rule Retriever is implemented based on the LegalBERT (Zhong et al., 2019). We set the learning rate, dropout rate, warmup steps, and max length of fact as 1×10^{-5}, 0.1, 200, and 500, respectively. The top-s is set from 1 to 4, and we report the best performance. The Rule Insight Generator, the Case Analyzer, and the Legal Reasoner are implemented by prompting LLMs. Moreover, we also provide the fine-tuned version of Legal Reasoner. For LoRA (Hu et al., 2022), we set the rank, α, and dropout rate as 16, 64, and 0.1, respectively. The maximum sentence length, maximum input sentence length, batch size, and learning rate are

[3]https://platform.deepseek.com/

[4]https://huggingface.co/Qwen

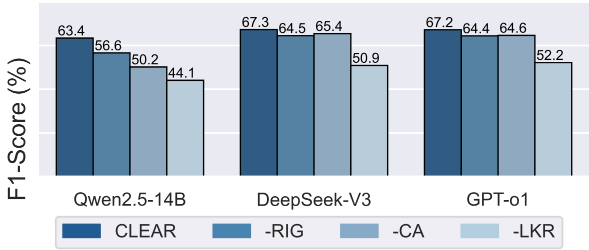

Figure 7: The results of the ablation study, where the Qwen2.5-14B, DeepSeek-V3, and GPT-o1 represent the backbone LLM of CLEAR.

set as 1560, 1024, 4, and 1×10^{-5}, respectively. All the experiments are conducted on 2 Tesla A100 40G GPUs with AdamW (Loshchilov and Hutter, 2019) optimizer for 10 epochs. We implement the baseline models based on the released source codes. The templates of prompt-based methods are detailed in the Appendix D. For the RAG-based methods, we retrieve n (from 1 to 4) similar cases based on the dense and sparse retriever (detailed in Appendix B), and we report the best performance in terms of the F1 score.

5.4 Main Results

Table 2 shows the overall performance. **(1)** It is observed that our CLEAR achieves superior performance, which is mainly attributed to the effectiveness of distinguishing confusing legal rules. For example, CLEAR based on the DeepSeek-V3 achieves 4.71%, 5.83%, and 7.34% improvements in terms of F1 score, recall, and precision, respectively. For the smaller backbone model Qwen2.5-14B, CLEAR surpasses the previous SOTA method by 8.26% in terms of precision, which highlights the efficiency of our framework. **(2)** We observe that CLEAR-SFT further enhances performance in discerning confusing legal rules, underscoring the high quality of the legal insights generated by our framework. **(3)** It is observed that the RAG-based methods achieve worse performance, indicating that similar cases may mislead the LLM.

5.5 Ablation Study

To validate the effectiveness of CLEAR's components, we conduct an ablation study and take the accuracy, macro precision, macro recall, and the macro F1 score as metrics. Figure 7 shows that performance decreases when removing the *Legal Knowledge Retriever* (-LKR), the *Rule Insight Generator* (-RIG) and the *Case Analyzer* (-CA), which demonstrates the contribution of each component of CLEAR. Specifically, for the -RIG, we remove the *Rule Insight Generator* module, and the *Case Analyzer* generates case insights based on the candidate legal paragraphs. The decline in performance indicates that the rule insights contribute to guiding LLMs in analyzing legal cases. For the -CA, we remove the Case Analyzer and ask LLM to choose the best answer from the candidate paragraphs based on the output of the *Rule Insight Generator*. The results show that removing the details of the legal case leads to a performance decline. For the -LKR, instead of fine-tuning LegalBERT (Zhong et al., 2019), we prompt LLMs to provide the most relevant paragraphs. It is observed that the performance has a sharp decline, indicating the effectiveness of our *Legal Knowledge Retriever*.

LKR	Top-1		Top-2		Top-3	
	Recall	AOR	Recall	AOR	Recall	AOR
LegalBERT	64.02	-	95.80	91.03	97.62	78.54
DeepSeek-V3	51.42	-	77.45	73.95	80.67	67.78
GPT-o1	49.69	-	75.32	71.50	83.32	68.89

Table 3: The recall rate (%) and AOR (%) of the LKR module. The AOR represents the rate of all the predicted paragraphs that belong to the ground truth law article.

Moreover, Table 3 shows the recall rate and Article Overlap Rate (AOR) of the LKR module, where the AOR denotes the proportion of all the recalled paragraphs under the ground truth legal article. The LKR module is implemented based on different LMs such as DeepSeek (DeepSeek-AI et al., 2025) and LegalBERT (Zhong et al., 2019). Take the top-2 predicted paragraphs as an example, we observe a high recall rate, and most of the predicted paragraphs belong to the ground truth legal article. This observation indicates that both LLMs and small language models struggle to distinguish legal paragraphs within the same article.

5.6 In-Depth Analysis

To evaluate the efficacy of our CLEAR, we attempt to answer the following research questions: **RQ-1: Paragraph Memorization**. Are LLMs capable of memorizing legal paragraphs? **RQ-2: Paragraph Distinguishing**. Does LLM know the differences between confusing paragraphs? **RQ-3: Paragraph Application**. How do LLMs perform on legal paragraph reasoning in the real-world judicial scenario?

Paragraph Memorization (RQ-1). The paragraph memorization task aims to assess whether LLMs possess the fundamental ability to retain and recall paragraph definitions. Unlike the previous study (Fei et al., 2024c), we focus on the more nuanced knowledge related to paragraphs. Specifically, we instruct LLMs to generate paragraph

Test Case		***Criminal Facts***: *In 2014,..., The defendant Wu, who was the party branch secretary of xxx Village at the time, used his position to collude with others to falsely report the land area and defraud the state compensation. The defendant Wu colluded with Wang and others to embezzle 62,323.10 yuan of compensation for the land occupation of the levee. In the spring of 2014, part of the reclaimed land behind the Heilongjiang levee section of Wang* ***Paragraph (Ground Truth)*** : ***Article 382 Paragraph 2***
Few-shot COT		*The defendant Wu, ..., In the process of handling matters involving state funds (i.e. compensation for land occupied by the levee), he took advantage of his position and colluded with others to falsely report the area of land occupied to defraud state compensation...,. In summary, the applicable paragraph may be the* ***Article 383 Paragraph 1***.
Hybrid Approach (Hierarchy)		*Candidate paragraphs: A. Article 382 Paragraph 2 T; B. Article 382 Paragraph 1 T.* *According to the facts and legal provisions provided, we infer the possible applicable article may be Choice B*
Hybrid Approach (Predicted)		*Candidate paragraphs: A. Article 382 Paragraph 1 T; B. Article 382 Paragraph 2 T.* *According to the facts and legal provisions provided, we infer the possible applicable paragraph may be Choice A*
RAG	Retrieved Case	***Criminal Facts***: *.... The defendant Wei, as the village chief of xxx Village, was entrusted by the town government to solicit opinions from the villagers of xxx Village, and signed compensation agreements and issued compensation money with the villagers who agreed to return the old homesteads and the villagers whose crops were cut down. ...* ***Paragraph (Ground Truth)*** *: Article 382 Paragraph 2*
	Answer	*Based on the facts provided, Wei defraud public funds for personal living expenses we infer the possible applicable paragraph may be Article 382 Paragraph 1 .*
CLEAR (Ours)	**LKR**	*A. Article 382 Paragraph 1 T; B. Article 382 Paragraph 2 T.*
	RIG	*Is the perpetrator a person entrusted by a state agency, state-owned company, enterprise, institution, or people's organization to manage and operate state-owned property?*
	CA	*The defendant Wu, as the then Party Secretary of xxx Village, was a person entrusted by the state organ... ... met the identification criteria of* ***"personnel entrusted to manage and operate state-owned property" in the Criminal Law, and his status fell within the category of state employees performing public duties in accordance with the law***
	LR	*The applicable legal provisions is* ***Article 382 Paragraph 2***

Figure 8: An example of the pipeline of retrieving legal knowledge (LKR), generating rule insights (RIG), analyzing case (CA) and legal reasoning (LR) by our CLEAR and comparing with other baseline models, where the T represents the definition of the paragraph.

LLM	Rouge-L (%)			Professionality Score	
	Recall	Precision	F1	PM	PD
Qwen2.5-14B	77.12	**16.78**	26.11	4.51	3.25
DeepSeek-V3	**89.57**	15.33	24.94	**4.77**	**3.97**
GPT-o1	89.55	16.44	**26.50**	4.75	3.85

Table 4: The results of paragraph memorization distinguishing evaluation, where the best results are marked in **bold**. The PM and PD represent the paragraph memorization and paragraph distinguishing, respectively.

definitions by supplying the legal article code and paragraph identifier (*e.g.*, *Please output the definition of the Law article 234 Paragraph 1 according to the Chinese Criminal Law*) and employ three legal experts to evaluate the generated text. Following previous studies (Fei et al., 2024c; Xu et al., 2024b), we take the Rouge-L and legal professionality as metrics, where the profissionality score (defined in Appendix C) ranges from 1 (lowest) to 5 (highest). As shown in Table 4, it is observed that LLMs achieve promising results in terms of the recall rate of Rouge-L and the professionality score (PM), demonstrating that LLMs are equipped with the basic legal knowledge. LLMs have a performance drop in precision and F1 score, which may be due to LLMs generating external content when they respond to the instruction.

Paragraph Distinguishing (RQ-2). In our CLEAR, given the candidate paragraphs, the rule insight generator raises questions that will be used to guide LLMs to analyze the legal case. The quality of generated questions significantly influences the output of subsequent modules in our CLEAR. In this section, to evaluate the professionality of the questions raised by the rule insight generator module from a legal perspective, we randomly sample 200 cases and employ three legal experts to perform a human evaluation. The professionality score ranges from 1 (lowest) to 5 (highest), which is defined in Appendix C. The evaluation results are shown in the last column of Table 4. It is observed that LLMs know the key differences between paragraphs but struggle to cover all the differential elements, indicating that there is a lot of room for improvement. Please refer to the Appendix C for more evaluation details and examples.

Paragraph Application (RQ-3). The case analyzer generates insight by answering the acquired questions, guiding LLMs to focus on key information and substantially reducing their exposure to distracting and irrelevant content (Deng et al., 2024). Figure 8 shows an example of the case insights generated by our CLEAR with the guidance of the questions (Please refer to Appendix C for more examples). Moreover, we also present the reasoning process of strong baselines. It is observed that the case analyzer can help LLMs to eliminate non-essential information in the context, guiding LLMs to analyze the case by focusing on important details. This strategy contributes to the improvement of LLMs' capability for discerning confusing legal paragraphs.

6 Conclusion

In this study, we introduce a novel framework CLEAR, containing four components: the legal rules retriever module aims to retrieve candidate legal paragraphs. The rule insight generator aims to prompt LLMs to raise questions to distinguish them. The case analyzer aims to generate case insights by answering the questions. Finally, the legal reasoner concludes by integrating legal knowledge and legal insights. Comprehensive experiment results demonstrate the effectiveness of our CLEAR.

Limitations

In this work, we propose a novel framework CLEAR aiming to distinguish confusing legal paragraphs. The limitation of the work is that CLEAR discerns the confusing legal rules with the guidance of the legal rule insights to analyze the legal case. Although promising results are obtained, the framework merely relies on their definitions. However, judges also consider the precedents and public opinion in the real scenario. In future work, we will explore how to integrate these factors into LLMs.

Ethics Statement

Each case in our benchmark LPP sources from the publicly available government website China Judgment Online [5] (CJO). Sensitive information, such as the license plate number, is anonymized to protect privacy. It is notable that our work aims to serve as a tool to help legal professionals distinguish confusing legal rules. Like many other models, CLEAR may generate some uncontrollable content during the legal reasoning process. Thus, we underscore that our framework is intended solely as a supplementary instrument within the legal domain, with final decision-making authority invariably resting with qualified legal professionals.

References

Zhenwei An, Quzhe Huang, Cong Jiang, Yansong Feng, and Dongyan Zhao. 2022. Do charge prediction models learn legal theory? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3757–3768.

BaiChuan-Inc. 2023. A large-scale 7b pretraining language model developed by baichuan-inc.

[5]https://wenshu.court.gov.cn/

Ryan C. Barron, Maksim E. Eren, Olga M. Serafimova, Cynthia Matuszek, and Boian S. Alexandrov. 2025. Bridging legal knowledge and ai: Retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization.

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. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

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

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

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li,

Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report.

Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. 2024. Enabling discriminative reasoning in llms for legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 784–796.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization?

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391.

Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024a. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing.

Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024b. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, et al. 2025. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the International Conference on Machine Learning*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024c. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962.

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.

Wanwei He, Jiabao Wen, Lei Zhang, Hao Cheng, Bowen Qin, Yunshui Li, Feng Jiang, Junying Chen, Benyou Wang, and Min Yang. 2023. Hanfei-1.0. `https://github.com/siat-nlp/HanFei`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR*.

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

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, et al. 2024. Leveraging large language models for learning complex legal concepts through storytelling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7194–7219.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Zhuo, Patrick Emerton, and Genevieve Grant. 2023. Can ChatGPT perform reasoning using the IRAC method in analyzing legal scenarios like a lawyer? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13900–13923.

Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharrer, Megan Ma, and Marzieh Nabi. 2025. Towards robust legal reasoning: Harnessing logical llms in law.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Haitao Li. 2023. Lexilaw: A large-scale 6b pretraining language model in legal domain.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023. Xiezhi chinese law large language model. `https://github.com/LiuHC0428/LAW_GPT`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR*.

Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Prototype-based interpretability for legal citation prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4883–4898.

Kangcheng Luo, Quzhe Huang, Cong Jiang, and Yansong Feng. 2025. Automating legal concept interpretation with llms: Retrieval, generation, and evaluation.

Yougang Lyu, Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang, and Zhaochun Ren. 2023. Multi-defendant legal judgment prediction via hierarchical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2198–2209.

Neil MacCormick. 2005. Rhetoric and the rule of law: a theory of legal reasoning. In *OUP Oxford*.

Venkatesh Mishra, Bimsara Pathiraja, Mihir Parmar, Sat Chidananda, Jayanth Srinivasa, Gaowen Liu, Ali Payani, and Chitta Baral. 2025. Investigating the shortcomings of llms in step-by-step legal reasoning.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

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.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Xiao Wei, Qi Xu, Hang Yu, Qian Liu, and Cambria Erik. 2024. Through the mud: A multi-defendant charge prediction benchmark with linked crime elements. In *Proceedings of the 2024 Conference of the Association for Computational Linguistics: ACL 2024*.

Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024a. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024b. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, pages 53366–53397.

Yang Wu, Chenghao Wang, Ece Gumusel, and Xiaozhong Liu. 2024c. Knowledge-infused legal wisdom: Navigating llm consultation through the lens of diagnostics and positive-unlabeled reinforcement learning.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *arXiv preprint arXiv:2105.03887*.

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*.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024a. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL*, pages 13326–13365.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3086–3095.

Qi Xu, Xiao Wei, Hang Yu, Qian Liu, and Hao Fei. 2024b. Divide and conquer: Legal concept-guided criminal court view generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3395–3410.

Rujing Yao, Yiquan Wu, Tong Zhang, Xuhui Zhang, Yuting Huang, Yang Wu, Jiayin Yang, Changlong Sun, Fang Wang, and Xiaozhong Liu. 2025. Intelligent legal assistant: An interactive clarification system for legal question answering.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7577–7597.

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*, page 973–982.

Linan Yue, Qi Liu, Lili Zhao, Li Wang, Weibo Gao, and Yanqing An. 2024. Event grounded criminal court view generation with cooperative (large) language models. *CoRR*, abs/2404.07001.

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

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 9701–9708.

Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Open chinese language pre-trained model zoo. Technical report.

Huilin Zhong, Junsheng Zhou, Weiguang Qu, Yunfei Long, and Yanhui Gu. 2020b. An element-aware multi-representation model for law article prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6663–6668.

A Details of Benchmark LPP

To support our study, we construct the benchmark LPP to better evaluate the LLM's ability in legal reasoning. As shown in Figure 10, each instance from LPP contains the defendant's name, background information, criminal facts, law article, and the paragraph. Figure 9 shows that the law articles and paragraphs have a long-tail distribution, which is in line with the real-world scenario (Xiao et al., 2018). Moreover, we built the legal rule knowledge base containing legal rules' definitions in law. Table 6 shows several examples.

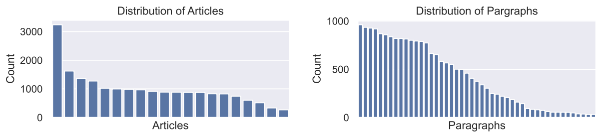

Figure 9: The long-tail distribution of dataset LPP.

An Example of the Case in LPP

id:4423288

dfd:["Tang"]

bg_info:{"Tang": The defendant Tang ... He was detained on April 16, 2017 for this case, and was sentenced to administrative detention for ten days on the 17th and 25th of the same month, and was released}

fact: the defendant Tang had a dispute with the victim Su and others over a property right dispute .... During the period, the defendant Tang beat the victim Su, ...

article: 234

paragraph: 1

Figure 10: An example of the case in LPP, where the dfd, bg_info, and fact represent the defendants related to the case, background information of the defendant, and the criminal facts, respectively.

B Retrieval-Augmented Generation

In our study, we adopt the Retrieval-Augmented Generation (RAG) as the baseline model. Following the previous study (Shui et al., 2023), the BM25 algorithm [6] is used to measure the similarity between the cases based on the criminal facts description, and we keep the top-k (from 1 to 4) cases with the highest similarity cases as demonstrations for LLMs. Moreover, following this strategy, we also implement the retriever by leveraging sentence-

[6]https://pypi.org/project/rank-bm25/

Score	Legal Professionality (Paragraph Memorization)	Legal Professionality (Paragraph Distinguishing)
1	No Professionality. The generated texts are unrelated to the article or paragraph.	No Professionality. The generated texts are unrelated to the article or paragraph.
2	Almost Professional. The generated texts are related to other articles and paragraphs.	Almost Professional. The generated texts are related to the paragraphs but without pointing out their difference.
3	Semi-Professionality. The generated texts are related to the article but not the paragraph.	Semi-Professionality. The generated texts are partially related to differential legal elements.
4	Highly Professional. The generated texts are related to most of the paragraph.	Highly Professional. The generated texts encompass the key distinguishing legal components.
5	Exactly Professional. The generated texts are related to the paragraph.	Exactly Professional. The generated texts exhibit complete coverage of all differential legal elements.

Table 5: The criteria of evaluation for the research questions of paragraph memorization and distinguishing.

Article	Paragraph	Legal Definition
234	1	Whoever intentionally injures another person shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention, or public surveillance.
234	2	Whoever intentionally injures another person causing serious bodily harm shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years
236	1	Anyone who rapes a woman by means of violence, coercion or other means shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years.
236	2	Anyone who rapes a girl under the age of fourteen shall be treated as rapist and punished more severely.
236	3	Whoever rapes a woman or a young girl shall be sentenced to fixed-term imprisonment of not less than ten years, life imprisonment or death if any of the following circumstances exist: (1) the circumstances of the rape are aggravated; (2) the rape of a woman or a young girl is committed against multiple persons; (3) the rape of a woman or a young girl is committed in public; (4) the rape is committed by two or more persons; (5) the rape is committed against a girl under the age of ten or causes injury to a girl; (6) the rape causes serious injury, death or other serious consequences to the victim.
...	...	....

Table 6: The examples of legal rule definition.

BERT [7], SimCSE [8], and ChatLaw [9] to measure the similarity between cases. Table 7 shows the overall performance. Finally, we take the best performance in terms of F1-score as the baseline performance.

[7]https://sbert.net/

[8]https://huggingface.co/cyclone/simcse-chinese-roberta-wwm-ext

[9]https://huggingface.co/chestnutlzj/ChatLaw-Text2Vec

Retriever	BackBone LLM	Metrics(%)			
		Accuracy	Precision	Recall	F1-Score
BM-25	Qwen2.5-14B	8.85	8.05	8.62	8.21
	Lexi-Law	9.60	9.88	10.46	10.92
	DISC-Law	9.44	10.40	8.15	9.01
	DeepSeek-V3	48.28	46.15	47.31	46.01
	CPT-o1	47.45	44.50	48.34	47.33
BERT	Qwen2.5-14B	11.30	10.54	9.20	9.33
	Lexi-Law	9.51	10.29	9.11	10.12
	DISC-Law	10.18	9.97	9.91	9.85
	DeepSeek-V3	49.38	47.31	45.61	45.76
	CPT-o1	48.40	45.87	49.57	47.78
SimCSE	Qwen2.5-14B	15.46	15.12	12.54	13.46
	Lexi-Law	13.30	11.43	9.022	9.85
	DISC-Law	15.80	10.32	8.22	9.67
	DeepSeek-V3	46.23	43.56	45.45	43.90
	CPT-o1	50.21	44.76	45.52	43.20
ChatLaw	Qwen2.5-14B	17.84	14.32	16.78	14.71
	Lexi-Law	15.60	12.88	9.46	10.75
	DISC-Law	12.46	10.76	8.93	8.83
	DeepSeek-V3	49.50	44.33	45.90	45.05
	CPT-o1	49.65	47.33	47.57	46.50

Table 7: The results of RAG-based methods.

C In-Depth Analysis

To evaluate the LLMs' ability of memorizing and distinguishing legal paragraphs, following previous studies (Fei et al., 2024c; Xu et al., 2024b), the Rouge-L, and professionality score are used as the metrics. Table 5 shows the definitions of the Professionality score. For the evaluation of paragraph distinguishing, we employ 3 legal experts to conduct the evaluation process on 300 randomly selected cases. Specifically, first, we ask legal experts to assign professionality score for cases selected base on definitions in Table 5. Each case is assigned professionaliy score three times independently. Secondly, cases with inconsistent scores are re-evaluated until a unanimous agreement is achieved. Finally, we report the average professionality score of the selected cases. Table 8 shows several examples of the evaluation results.

Paragraphs	**Article 234, Paragraph 1**: "Whoever intentionally injures another person shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance"; **Article 264, Paragraph 1**: "Whoever steals a large amount of public or private property, or steals repeatedly, breaks into a house, steals with a weapon, or pickpockets, shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance, and shall also or solely be fined"
Qwen2.5-14B	Is the behavior of the suspect an infringement of property or personal rights? ##**Score:4**##
DeepSeek-v3	Is the subjective purpose of the perpetrator to harm others or to illegally occupy public or private property? ##**Score:4**##
GPT-o1	Did the suspect commit theft or intentional injury to others? ##**Score:4**##
Paragraphs	**Article 236, Paragraph 1**: Anyone who rapes a woman by violence, coercion or other means shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years. **Article 236, Paragraph 3**: Anyone who rapes a woman or a young girl shall be sentenced to fixed-term imprisonment of not less than ten years, life imprisonment or death if any of the following circumstances exist: (1) The rape of a woman or a young girl is of a serious nature; (2) The rape of a woman or a young girl is committed against multiple persons; (3) The rape of a woman or a young girl is committed in public; (4) The rape is committed by two or more persons; (5) The rape of a girl under the age of ten or causes injury to the girl; (6) The rape causes serious injury, death or other serious consequences to the victim.
Qwen2.5-14B	Does the criminal act have the following circumstances: aggravated circumstances, multiple victims, public places, gang rape, involving girls under ten years old, or causing serious consequences? ##**Score:4**##
DeepSeek-v3	Does the criminal act meet any of the following circumstances: rape of women or girls with aggravated circumstances; rape of women or girls with multiple victims; rape of women or girls in public places; gang rape by more than two persons; rape of girls under ten years old or causing harm to girls; causing serious injury, death, or other serious consequences to the victim? ##**Score:5**##
GPT-o1	Does the criminal act have one of the six aggravating circumstances listed in Article 236, paragraph 3? ", "Is the victim under fourteen years old? ", "Is the victim under ten years old or has the girl been harmed? ##**Score:4**##
Paragraphs	**Article 236, Paragraph 1**: Anyone who rapes a woman by violence, coercion or other means shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years. **Article 236, Paragraph 2**: Anyone who rapes a girl under the age of fourteen shall be deemed to have raped her and shall be punished more severely.
Qwen2.5-14B	Is the victim under the age of 14? ##**Score:5**##
DeepSeek-v3	Is the victim over the age of 14? ##**Score:5**##

GPT-o1	Is the victim over the age of 14?##**Score:5**##

Table 8: Examples of differences between the confusing legal paragraphs generated by LLMs.

D Prompt Template of Baselines

Table 9 shows the prompt template of baselines methods containing the zero/few-shot methods, the similar case retriever-enhanced method and the hybrid approach.

Baseline	Prompt Template
Zero-Shot	You are a helpful legal assistant, please predict the applicable paragraph, for example the Article 234 Paragraph 1. Criminal Fact: {Criminal Fact of Input Case}.
Few-Shot	You are a helpful legal assistant, please predict the applicable paragraph, for example the article 234 paragraph 1. Criminal Fact: {Criminal Fact of Demonstration} Paragraph: {The Article xxx Paragraph xxx} Criminal Fact: {Criminal Fact of Input Case}
Similar Case Retriever Enhanced	You are a helpful legal assistant, please predict the applicable paragraph, for example the article 234 paragraph 1. Criminal Fact: {Criminal Fact of Retrieved Similar Case} Paragraph: {The Article xxx Paragraph xxx} Criminal Fact: {Criminal Fact of Input Case}.
Hybrid Approach	You are a helpful legal assistant, please select the applicable paragraph from the candidate. Criminal Fact: {Criminal Fact of Input Case} Candidate Paragraph: {The Article xxx Paragraph xxx; The Article xxx Paragraph xxx; The Article xxx Paragraph xxx; The Article xxx Paragraph xxx}

Table 9: Prompt Template of Baseline Methods.

E Case Analysis

Table 10 shows several examples of the decision pipeline of our CLEAR. We can observe that our framework can distinguish the confusing paragraphs with the guidance of the questions raised by the rule insights generator and case analysis results.

Case	**Criminal Facts:** At about 10:00..., When police officer Zhang was directing the auxiliary police officers on the scene to control the defendants Lu Jingui and Zhang Shugen in accordance with the law, the defendant Lu Jingui bit the auxiliary police officer Zou with his mouth and kicked him in the thigh, and scratched the auxiliary police officer Zheng's arm with his nails; the defendant Zhang Shugen burned the face of auxiliary police officer Du with a cigarette.... **Ground Truth:** Article 277 Paragraph 5

Legal Knowledge Retriever	**Article 277 Paragraph 1:** "Whoever uses violence or threats to obstruct state organ staff from performing their duties according to law shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention, controlled release or a fine."**Article 277 Paragraph 5:** "Whoever violently attacks a people's police officer who is performing his duties in accordance with the law shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance; whoever uses firearms, controlled knives, or drives a motor vehicle to hit a police officer, thereby seriously endangering his personal safety, shall be sentenced to fixed-term imprisonment of not less than three years but not more than seven years."
Rule Insights Generator	Were the victims of the violent attack people's police who were performing their duties in accordance with the law?
Case Analyzer	The targets of the violent attack were the people's police who were performing their duties in accordance with the law. According to the case description, the auxiliary police officers Zou, Du and Zheng were attacked violently when they were controlling the defendants Lu Jingui and Zhang Shugen in accordance with the law under the command of the police officer Zhang, and the auxiliary police performed their duties under the command of the police, which was an act of performing their duties in accordance with the law. Therefore, the targets of the violent attack were the people's police who were performing their duties in accordance with the law.
Legal Reasoner	The applicable legal provisions is **Article 277 Paragraph 5**
Case	**Criminal Facts:** In 2012...,Wang Wanlong was the village accountant at the time and was responsible for handling the money..., In order to achieve the purpose of embezzling 10,000 yuan, one day in the first half of 2014, Wang Wanlong asked Pang Mou 1 to write the receipt of 20,000 yuan as 30,000 yuan, and took the receipt of 30,000 yuan to the village director Pang Mou 2 for signature and then reported it to the village finance department. The stolen money was used for family expenses... **Ground Truth:** Article 382 Paragraph 2
Legal Knowledge Retriever	**Article 382 Paragraph 1:** "If a state functionary takes advantage of his or her position to embezzle, steal, defraud or illegally possess public property by other means, it is a crime of corruption." **Article 382 Paragraph 2**: "Any person who is entrusted by a state organ, state-owned company, enterprise, institution or people's organization to manage or operate state-owned property and who uses his or her position to embezzle, steal, defraud or illegally occupy state-owned property by other means shall be considered to have committed corruption."
Rule Insights Generator	Is the perpetrator a person who is entrusted by a state organ, state-owned company, enterprise, institution or people's organization to manage and operate state-owned property?

Case Analyzer	The perpetrator, Wang Wanlong, is a person who is entrusted by state organs, state-owned companies, enterprises, institutions, and people's organizations to manage and operate state-owned property. According to the case description, Wang Wanlong was the village accountant at the time and handled the "one issue, one discussion" project reward and subsidy funds of 60,948 yuan obtained from the county finance bureau, which shows that he was entrusted to manage state-owned property. In addition, he took advantage of his position and embezzled 10,000 yuan of public funds through false means, further proving that he is a person who is entrusted to manage and operate state-owned property.
Legal Reasoner	The applicable legal provisions is **Article 382 Paragraph 2**

Table 10: Examples of distinguishing confusing legal rules process of our framework CLEAR.