Can Large Language Models Grasp Legal Theories? Enhance Legal Reasoning with Insights from Multi-Agent Collaboration

Weikang Yuan[1,2], Junjie Cao[2], Zhuoren Jiang[1] [*], Yangyang Kang[1], Jun Lin[2], Kaisong Song[3,2], Tianqianjin Lin[1,2], Pengwei Yan[1,2], Changlong Sun[2], Xiaozhong Liu[4]

[1]Zhejiang University, [2]Tongyi Lab, Alibaba Group
[3]Northeastern University, [4]Worcester Polytechnic Institute

{yuanwk, jiangzhuoren, yangyangkang, lintqj, yanpw}@zju.edu.cn, {junjie.junjiecao, linjun.lj, kaisong.sks}@alibaba-inc.com, changlong.scl@taobao.com, xliu14@wpi.edu

Abstract

Large Language Models (LLMs) could struggle to fully understand legal theories and perform complex legal reasoning tasks. In this study, we introduce a challenging task (confusing charge prediction) to better evaluate LLMs' understanding of legal theories and reasoning capabilities. We also propose a novel framework: Multi-Agent framework for improving complex Legal Reasoning capability (MALR). MALR employs non-parametric learning, encouraging LLMs to automatically decompose complex legal tasks and mimic human learning process to extract insights from legal rules, helping LLMs better understand legal theories and enhance their legal reasoning abilities. Extensive experiments on multiple real-world datasets demonstrate that the proposed framework effectively addresses complex reasoning issues in practical scenarios, paving the way for more reliable applications in the legal domain.

1 Introduction

Large Language Models (LLMs) have shown remarkable generalization ability across diverse range of tasks and applications (Chowdhery et al., 2023; Touvron et al., 2023; OpenAI, 2023). But, current benchmarks may not adequately reflect the reasoning capabilities of LLMs (Valmeekam et al., 2024) and do not accurately reflect real-world situations (Huang and Chang, 2023). The validation of LLMs in more realistic and meaningful applications, such as legal reasoning, still requires extensive exploration.

In the legal domain, the core competency of legal professionals is to apply legal rules to facts and draw conclusions, as described by the IRAC (Issue, Rule, Application, Conclusion) framework. As shown in Figure 2, a legal professional can determine whether a case fact conforms to specific criminal charges based on legal rules. They critically assess a case against potential charges, focusing on the key points of relevant legal rules, to accurately identify the appropriate charge and distinguish inapplicable charges. Legal rules, which manifest legal theories, determine the legal consequences of factual situations (MacCormick, 2005). Therefore, properly applying legal rules reflects the grasp of legal theories.

However, powerful LLMs may struggle to fully understand legal theories and perform basic legal reasoning tasks. Existing study (Dahl et al., 2024) has found that when LLMs are given criminal facts and legal rules, then asked whether cases constitute a certain charge, they tend to answer "yes", regardless of whether the charge is correct (golden charge) or a closely related one (confusing charge). Our empirical experiments also confirmed this issue. We sampled real-world criminal cases involving the charge of **Misappropriation of Public Fund**, inputting the criminal facts and legal rules into LLMs, and asked whether the case constituted the golden charge. Meanwhile, we created a control group where we input the same criminal facts and related legal rules, asking whether the case constituted a confusing charge (**Fund Misappropriation**). These two charges are very similar, with the key difference being *whether the defendant's subject position is that of a state functionary*. As shown in Figure 1, when performing legal reasoning, regardless of the prompt method or the version of GPT used, LLMs exhibit significant declines in performance when predicting confusing charges.

Generally, LLMs could face following challenges in legal reasoning: **Inconsistent reasoning.** Legal reasoning involves multi-step, compositional logic processes (Servantez et al., 2024). LLMs can be easily distracted by the interaction when generating reasoning steps (Shi et al., 2023) and may not be trustworthy by the tendency to give affirmative answers (Dahl et al., 2024). **Missing key details.** Legal rules and criminal facts are often

[*]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7577–7597
November 12-16, 2024 ©2024 Association for Computational Linguistics

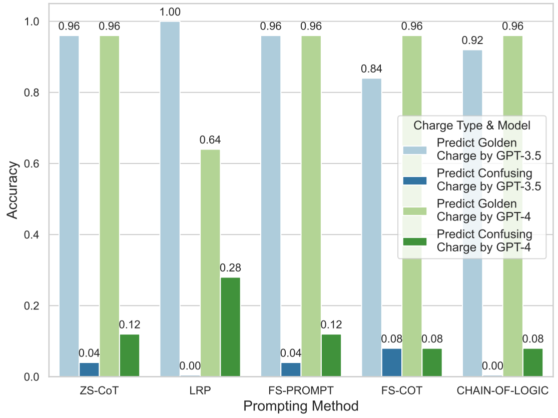


Figure 1: The performance of LLMs on predicting the golden (**Misappropriation of Public Fund**) or confusing charge (**Fund Misappropriation**) for the cases from CAIL-2018 datasets. The horizontal axis represents 5 advanced promt methods to solve legal reasoning problems (detailed information is described in Section 5). In each method, GPT-3.5 and GPT-4 both exhibit a significant performance gap.

described in complex natural language, making it challenging for LLMs to fully understand and reason based on them. Consequently, they often overlook key information in the rules. **Lacking domain knowledge.** LLMs may hallucinate erroneous legal knowledge (Dahl et al., 2024) or encounter gaps in common-sense knowledge (Huang et al., 2023). Their overconfidence can obscure these shortcomings, making them difficult to identify (Ni et al., 2024).

To better evaluate LLMs' understanding of legal theories and their reasoning capabilities, we introduce and construct a challenging task: confusing charge prediction (The detailed task definition is provided in Section 3). We also propose a novel framework: Multi-Agent framework for improving complex Legal Reasoning capability (MALR). First, an auto-planner breaks down complex legal rules into sub-tasks, allocating them to expert agents, reducing inconsistent reasoning in LLMs. Second, a non-parametric learning framework is proposed to draw adaptive rule-insights from trials and errors. To address the problem that LLMs may overlook crucial information in legal rules, we design a module that mimics human learning by gaining experience through reasoning trajectories and knowledge feedback, then learning insights through self-reflection. These insights supplement the rules, encouraging LLMs to focus on key factors from legal knowledge and fully understand the rules, while also guiding them to automatically seek help when they feel uncertain. These designs effectively improve LLMs' reasoning and critical-thinking skills.

Our contributions are threefold:

- We propose a multi-agent framework based on non-parametric learning, which encourages LLMs to automatically decompose complex legal tasks and extract insights from legal rules. Our framework assists LLMs in gaining a deeper understanding of legal rules and enhances their legal reasoning capabilities.
- We introduce a challenging task, confusing charge prediction, to better evaluate LLMs' understanding of legal theories and their reasoning capabilities.
- Extensive experiments are conducted on the multiple real-world datasets, demonstrating that the proposed framework can effectively addresses complex reasoning issues in real-world scenarios. Our work paves the way for more trustworthy application in legal domain[1].

2 Related Work

2.1 Legal AI and LLMs

Legal AI aims to improve legal tasks through AI techniques, particularly showing significant potential in alleviating the issue of "too many cases but too fewer legal experts" in the legal field (Katz et al., 2023; Dahl et al., 2024). One of the main challenges in legal domain is the training dataset can be considerably expensive and sparse (Sun et al., 2020), primarily comes in text, such as statutes, law articles and criminal cases. Under these circumstances, LLMs shows promising prospects in legal scenarios due to their powerful generalization capabilities in understanding and generating text. These applications include areas such as legal summarization (Deroy et al., 2023), legal document retrieval (Sun et al., 2024), legal question answering (Louis et al., 2024) and legal judgment prediction (Yu et al., 2022; Wu et al., 2023; Servantez et al., 2024).

2.2 Legal Reasoning and LLMs

Reasoning based on judicial rules and case fact descriptions is a fundamental ability of legal professionals, reflecting their understanding and application of legal theories (Servantez et al., 2024). Pre-

[1] Our source code can be found at `https://github.com/yuanwk99/MALR`

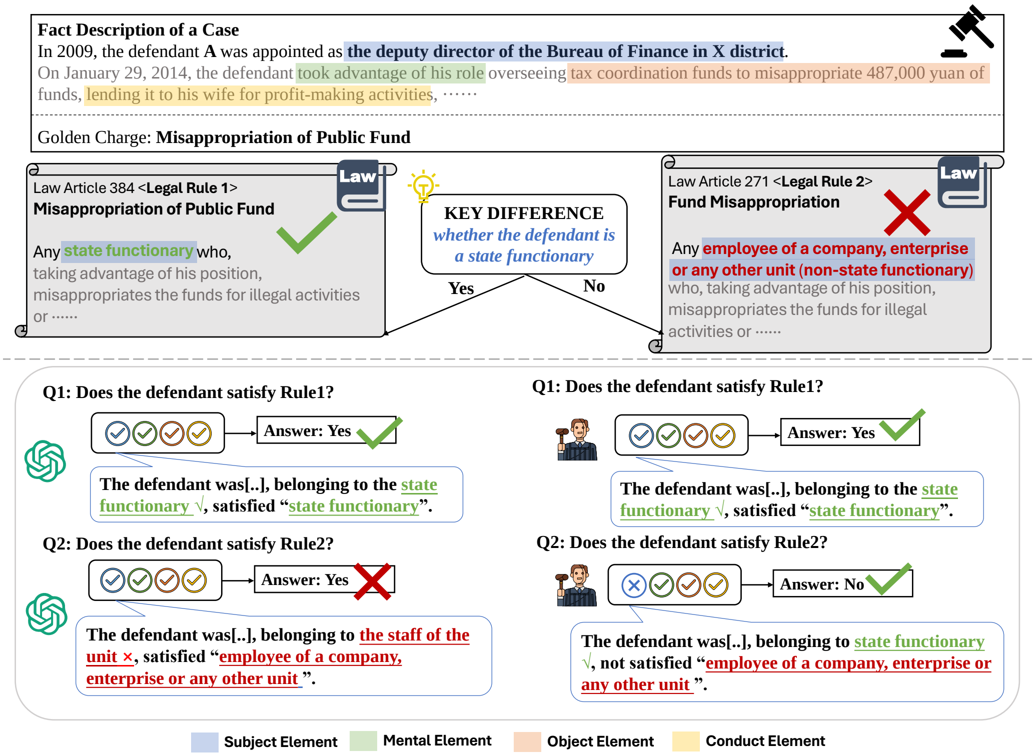

Fact Description of a Case
In 2009, the defendant A was appointed as the deputy director of the Bureau of Finance in X district.
On January 29, 2014, the defendant took advantage of his role overseeing tax coordination funds to misappropriate 487,000 yuan of funds, lending it to his wife for profit-making activities, ……
Golden Charge: Misappropriation of Public Fund
Law Article 384 <Legal Rule 1>
Misappropriation of Public Fund
Any state functionary who, taking advantage of his position, misappropriates the funds for illegal activities or ……
Law
KEY DIFFERENCE
whether the defendant is a state functionary
Yes
No
Law Article 271 <Legal Rule 2>
Fund Misappropriation
Any employee of a company, enterprise or any other unit (non-state functionary) who, taking advantage of his position, misappropriates the funds for illegal activities or ……
Law
Q1: Does the defendant satisfy Rule1?
Answer: Yes
The defendant was[..], belonging to the state functionary √, satisfied "state functionary".
Q2: Does the defendant satisfy Rule2?
Answer: Yes
The defendant was[..], belonging to the staff of the unit ×, satisfied "employee of a company, enterprise or any other unit ".
Q1: Does the defendant satisfy Rule1?
Answer: Yes
The defendant was[..], belonging to the state functionary √, satisfied "state functionary".
Q2: Does the defendant satisfy Rule2?
Answer: No
The defendant was[..], belonging to state functionary √, not satisfied "employee of a company, enterprise or any other unit ".
Subject Element
Mental Element
Object Element
Conduct Element

Figure 2: An example to demonstrate how a judge and an LLM apply legal rules to conclude whether a case satisfies a specific charge. This example outlines two confusing charges under Chinese criminal law: the **Crime of Fund Misappropriation** and the **Crime of Misappropriation of public fund**. The most significant difference between the two charges is whether the defendant is a state functionary. In the case description, the defendant is *"the Deputy Director of the Bureau of Finance in X district"*, a position that qualifies as a state functionary. Therefore, the judge can easily infer that the case falls under the Crime of Misappropriation of public funds, rather than Fund Misappropriation. However, the LLM fails to predict the confusing charge.

vious studies on Legal Judgment Prediction (LJP) have primarily focused on automatically predicting case charges mainly from fact descriptions (Zhong et al., 2018; Chalkidis et al., 2019; Liu et al., 2022). Additionally, similar precedents can be retrieved as supplementary guides to improve performance (Wu et al., 2023). However, this approach can lead to inaccurate judgments due to overlooked potential differences in case details. To address the subtle differences between case details and legal rules, knowledge graphs have been introduced to solve confusing charges problems (Yue et al., 2021; Li et al., 2024). Despite these efforts, utilizing Four Elements Theory and innocent datasets, An et al. (2022) found that charge prediction models do not take legal theories into consideration. Instead, models learn certain shortcuts for legal reasoning. Furthermore, Chain-of-Logic (Servantez et al., 2024) directly incorporates legal rules into prompts to elicit rule-based reasoning, achieving good performance on legal reasoning tasks involving three distinct rules from the LegalBench benchmark (Guha et al., 2024). SimuCourt proposes a multi-agent framework to simulate the decision-making process of a judge (He et al., 2024).

Unlike existing works, we aim to evaluate and enhance the capacity of LLMs to reason based on legal theories, rather than treating legal rules as supplementary information.

3 Preliminary

We propose **Confusing Charge Prediction Task** to evaluate the LLMs' ability to identify correct legal charges based on fact descriptions and legal rules. This task highlights subtle distinctions in legal rules. Only LLMs that capture these nuances can demonstrate their understanding of legal theory.

Fact Description: a concise description of a legal case, represented as a word sequence $\mathbf{f} = \{w_1, w_2, ..., w_l\}$. **Legal Rule**: the definition of a specific criminal charge from law articles, also a word sequence $\mathbf{r}_c = \{w_1, w_2, ..., w_n\}$, where c is the criminal charge. **Golden Charge**: The true crime label of a case. **Confusing Charge**: A charge similar to the golden charge but differing in one element (An et al., 2022).

To ensure LLMs' trustworthiness in applying legal rules, we require them to confirm the golden charge as True and reject the confusing charge as False. The task can be formalized as:

$$y = \Gamma(f, r_{gc}) \land \neg\Gamma(f, r_{cc})$$

where gc refers to the golden charge, cc refers to the confusing charge, and Γ is the charge prediction model. y is True only if the fact description f satisfies the rule of golden charge r_{gc} and does not match the rule of confusing charge r_{cc}.

LLMs should correctly identify the golden charge and explain why the fact description doesn't match the confusing charge, demonstrating understanding of legal theories.

4 The Proposed Framework

Figure 3 shows an overview of our proposed framework, which consists of four core components: Auto-Planner for Task Decomposition, Role Assignment for Sub-task Agent, Adaptive Rule-Insights Training, and Reasoning with Rule-Insights.

4.1 Auto-Planner

A single LLM may exhibit inconsistencies when directly generating the whole reasoning process (Wang et al., 2024). Therefore, we designed an automatic planning module to decompose the task. Given a question q, a case fact description $\mathbf{f}$, and the corresponding legal rule $\mathbf{r}_c$ about a criminal charge c, we guide an LLM as *auto-planner* to decompose the question into a sequence of sub-tasks based on the input of the fact and the rule:

$$[st_1, ..., st_k] = LLM(q, \mathbf{r}_c, \mathbf{f}, p_{auto}) \qquad (1)$$

where p_{auto} is the guideline prompt for LLMs to generate the sub-task set for the question q, and the st stands for the specific sub-task action and the k is the length of the sub-task set.

Given the resource-intensive nature of generating sub-tasks for every criminal case, we design a more effective strategy. We first sample a smaller scale dataset for auto-planner training, and prompt the LLM to generate the sub-task set for each sample. Subsequently, an LLM is used to identify semantically duplicate sub-task and compute the probability distribution for each sub-task. Based on this process, important sub-tasks with probability exceeding the threshold ζ are used to constitute the final sub-task set.

4.2 Assigning Roles to Sub-task Agent

Assigning roles can help the LLMs better perform complex task reasoning (Wang et al., 2023). Therefore, based on the sub-task set $[st_1, ..., st_k]$, we employ k LLM-based agents to tackle each subtask. Formally, we use the content of the sub-tasks to generate the appropriate prompts p_{st} and generate k agents to tackle each sub-task. Each agents will break down specific aspects of legal rule, check whether the fact description f conforms the corresponding sub-rule $r_{c_{st}}$ and generate the answer A_{st}. This process can be formulated as:

$$A_{st} = M_{st}(r_{c_{st}}, f, p_{st})$$

After obtaining the answer for each sub-task, a logic expression (Servantez et al., 2024) based on the principle of "presumption of innocence" (An et al., 2022) will generate the final answer.

4.3 Adaptive Rule-Insights Training

As aforementioned, LLMs can be easily distracted by the irrelevant context (Shi et al., 2023) and tend to overlook the key information and important details within rules. Therefore, we aim to enable LLMs to automatically extract the most critical information for legal judgement directly from the rules. Previous research demonstrated that LLMs can autonomously learn from their own experiences (Shinn et al., 2024; Zhao et al., 2024). Inspired by the Kolb's Experiential Learning Model (Kolb, 2014), we design the insights training module, as shown in Figure 3 (B), which consists of three core processes: experience gaining, insights drawing from errors and successes, and insights filtering. This module mimics the human learning process and facilitates the LLMs to automatically learn rules, discovering and summarizing the most critical information in the rules.

Experience Gaining. A small training dataset with N charges is constructed, with each charge containing case samples and corresponding confusing charges. Based on the fact descriptions of a case, sub-task agents M_{st} will respectively generate sub-answers for both golden charge and confusing charge. These sub-answers will be synthesized into a final determination of whether the case satisfies the legal rule for the golden charge or

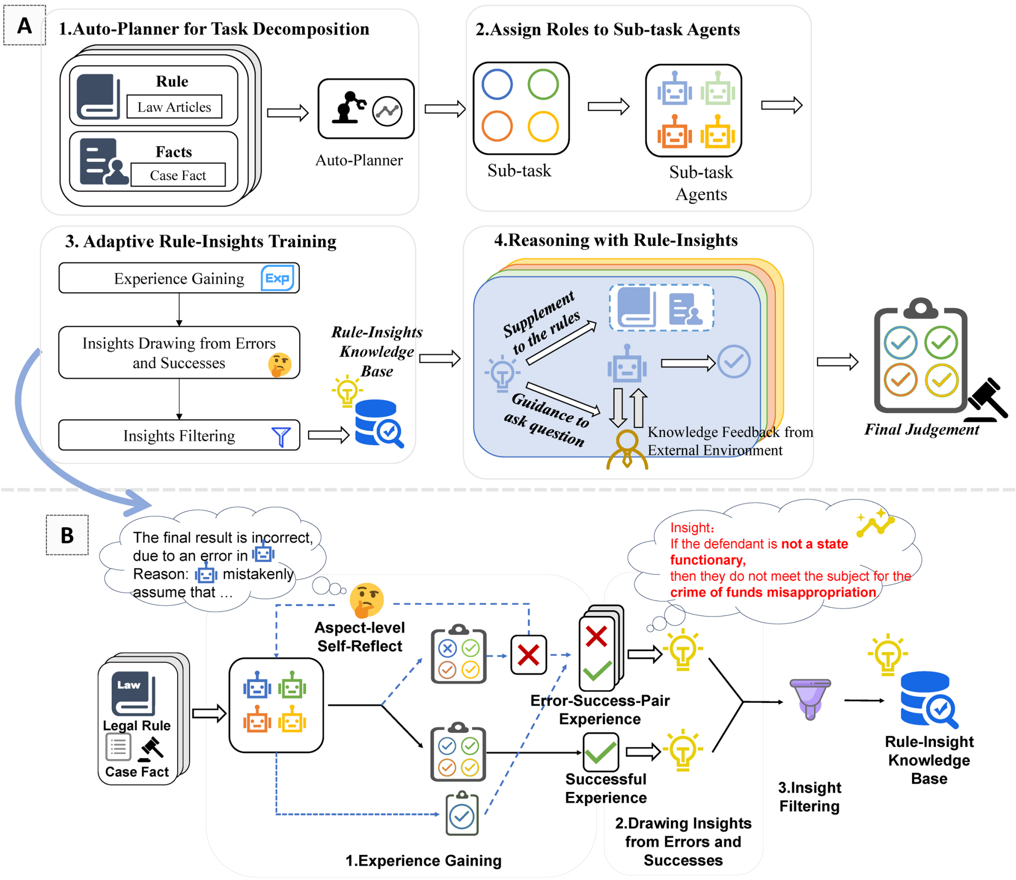

A
1.Auto-Planner for Task Decomposition
Rule
Law Articles
Facts
Case Fact
Auto-Planner
2.Assign Roles to Sub-task Agents
Sub-task
Sub-task Agents
3. Adaptive Rule-Insights Training
Experience Gaining
Exp
Insights Drawing from Errors and Successes
Insights Filtering
Rule-Insights Knowledge Base
4.Reasoning with Rule-Insights
Supplement to the rules
Guidance to ask question
Knowledge Feedback from External Environment
Final Judgement
B
The final result is incorrect, due to an error in
Reason: mistakenly assume that ...
Aspect-level Self-Reflect
Law
Legal Rule
Case Fact
Error-Success-Pair Experience
Successful Experience
1.Experience Gaining
2.Drawing Insights from Errors and Successes
3.Insight Filtering
Rule-Insight Knowledge Base
Insight:
If the defendant is not a state functionary,
then they do not meet the subject for the crime of funds misappropriation

Figure 3: Our research framework in (A) and Adaptive Rule-Insights training process in (B).

confusing charge, and the ground truth is used as feedback. Successful trials are saved as successful experience, while failed trials trigger the Aspect-level Self-Reflector to identify incorrect sub-task agents M^e_{st} and generate reasons rs_e for the errors. In the next trial, the error reasons are used to improve sub-task agents' predictions. Such approach of learning from trials and errors can be effective, as demonstrated in (Shinn et al., 2024). This iterative process continues for a maximum of L rounds, and corrected trajectories are retained as error-success-pair experience. The algorithm procession is detailed in Alg 1.

Insights Drawing from Errors and Successes. We gain insights into rules by analyzing experience collections using different methods. For error-success pairs, we use a contrast-based approach, comparing incorrect and correct attempts. This enables the $M_{insight}$ to identify the most critical task-level judgments where errors occur, outputting insights in an if-then format. Successful experiences reveal best practices (Zhao et al., 2024), so we provide the entire successful reasoning process to $M_{insight}$ to generate corresponding insights.

Insights Filtering. To address the potential for generating duplicate or incorrect insights when interpreting rules from the aforementioned process, we employ an LLM as an automatic checker, M_{filter}, to identify and remove redundant insights and filter out invalid ones. Ultimately, the retained insights are stored in the rule-insight knowledge base as memory. The pseudo-code for insight drawing and filtering is presented in Algorithm 2.

4.4 Reasoning through Insights

The generated insights serve two purposes: (1) they supplement the rules as additional notes, and (2) they guide LLMs to inquire about uncertainties when facing knowledge gaps in specific sub-tasks.

For implementation, we retrieve relevant insights in_{st} from the knowledge base I for each question to improve reasoning. If the rule does not exist, the most similar rule from the knowledge base is selected based on semantic similarity, and a few-shot method is used to generate new insights. To address potential knowledge gaps in LLMs, our in-

Alg 1: Experience Gaining

```
Initialze:Self-Reflector, Sub-task Agent,
  Evaluator: $M_{reflect}, M_{st}, M_e$
Number of charges $N$
Successful Experience $E_{success}$
Error-Success-Pair Experience $E_{esp}$
trajectory $\tau$, Fact Description $f$
Golden Charge $gc$, Confusing Charge $cc$
while charge $n \leq N$ do
    Set $r_{gc}, r_{cc} \leftarrow gc_n, cc_n$
    Generate $\tau_{l,gc} = [A_{1,gc}, ..A_{k,gc}]$ using
      $M_{st}, r_{gc}, f$
    Generate $\tau_{l,cc} = [A_{1,cc}, ..A_{k,cc}]$ using
      $M_{st}, r_{cc}, f$
    while trial $l \leq L$ do
        Evaluate $(\tau_{l,gc}, \tau_{l,cc})$ using $M_e$
        if success then
            if $l = 1$ then
                Append $\tau_{l,gc}, \tau_{l,cc}$ to
                  $E_{success}$
                break
            else
                Append $\tau_{l,gc}, \tau_{l,cc}$ to $E_{esp}$
                break
            end
        else
            Identify error $M_{st}^{e}$ and Generate
              $rs_e$ using $M_{reflect}$
            $e \in \{gc, cc\}$
            Generate $A_{k,e}$ using
              $M_{st}^{e}, r_e, f, rs_e$
            Updating $\tau_{l+1,e}$ using $A_{k,e}$
        end
    end
end
```

sights serve as guidance to ask factuality questions. Based on the insights, we identify sub-tasks requiring fact-checking and use a few-shot method to prompt LLMs to ask key questions like "Is a <job position> a <state functionary>?" The generated question is then presented to a knowledgeable expert (a legal professional, a domain-specific LLM, or a search engine) to obtain knowledge feedback kg_{st}. Finally, we incorporate related insights in_{st} and knowledge feedback kg_{st} into our ultimate reasoning process. As shown in Figure 3(A)(4), the improved reasoning process for each sub-task agent can be represented as:

$$A_{st} = M_{st}(r_{st}, f, in_{st}, kg_{st}, p_{st}) \tag{2}$$

Alg 2: Insight Drawing and Filtering

```
Initialze:Insight-Drawer, Insight-Filter:
  $M_{insight}, M_{filter}$
Successful Experience $E_{success}$
Error-Success-Pair Experience $E_{esp}$
Number of charges $N$
Number of sub-task $k$
Rule-Insight Knowledge Base $I$
while charge $n \leq N$ do
    Construct error-success pair of sub-task
      trial from $E_{esp}$:
      $\mathbf{P} = \{(A_{st_k}^{error}, A_{st_k}^{success}), ...\}$
    for each $p$ in $\mathcal{P}$ do
        Drawing insight $i$ using $M_{insight}(p)$
        Update $i$ to $I[charge][st_k]$
    end
    for each $exp$ in $E_{esp}$ do
        Drawing insight $i$ using
          $M_{insight}(exp)$
        Update $i$ to $I[charge][st_k]$
    end
    Filter $I[charge]$ using $M_{filter}$
end
```

All prompt templates for our MALR agents are provided in Appendix A.

5 Experiment

5.1 Experimental Setting

Dataset. We evaluate LLMs' legal reasoning capability on three datasets: CAIL2018 (Xiao et al., 2018), CJO (Wu et al., 2023), and CAIL-I (An et al., 2022). CAIL2018 and CJO consist of real-world cases with fact descriptions and golden charges. We match corresponding confusing charges based on the golden charges and randomly sample 400 cases from CAIL2018 and 100 from CJO for evaluation. CAIL-I's testset contains 462 innocent cases without crimes and the most similar charges to each non-criminal fact. Further dataset information is available in Appendix B.

The pairs of confusing charges are carefully selected by a group of legal experts, including: (1) Misappropriation of Public Fund (MP) v.s. Fund Misappropriation (FM); (2) Bribery (BY) v.s. Bribery of Non-State Officials (BN); (3) Kidnapping (KD) v.s. Illegal Detention (ID); (4) Fraudulently Obtaining Loans (FL) v.s. Loan Fraud (LF); (5) Fund Misappropriation (FM) v.s. Official Embezzlement (OE); (6) Fraud (FD) v.s. Loan Fraud

(LF); (7) Fraud (FD) v.s. Cheating and Bluffing (CB); (8) Forging, Altering, Trading Official Documents, Certificates and Seals of State Organs (FO) v.s. Forging the Seals of Companies, Enterprise, Institution (FS). Key differences between each pair are provided in Appendix B.

Implementation. We employ the publicly available GPT-3.5-Turbo-0125 and GPT-4-0125-preview models, with the temperature set to 0 for all text generation tasks. For auto-planner and insights training, we construct a small training set from CAIL-2018 training set. Specifically, we sample two cases for each of the 16 charges (totally 32 training samples). The threshold ζ for the sub-task auto-planner is set to 0.8. Sentence-BERT (Thakur et al., 2021) and cosine similarity are used to compute semantic distances between rules and unseen legal rules, facilitating rule-insight inference testing in CJO and CAIL-I. During the insights training period, we limit the maximum number of trial attempts L to 2. For providing knowledge feedback, we employ Farui-200B[2], which can be replaced by other legal models or even legal experts in real-world scenarios. Additionally, we construct a legal rule knowledge base that includes Chinese Criminal Law Articles and all charge definitions. All legal rules are retrieved from this knowledge base based on the charge name. The inference time and cost can be seen in Appendix B.

5.2 Baselines

Zero-shot Setting: (1) ZS-CoT (Kojima et al., 2022) uses "*Let's think step by step*" to encourages LLMs to generate intermediate steps and improve reasoning. (2) Legal Reasoning Prompting (LRP) (Yu et al., 2022) is a zero-shot legal prompting method that teaches LLMs to reason like a lawyer, following the "Approach, Issue, rule, application and conclusion" framework.

Few-shot Setting: (1) Few-Shot prompting (Brown et al., 2020) is the standard prompting method includes only the sample and answer. We use a two-shot setting with one positive and one negative examples. (2) Few-Shot CoT (Wei et al., 2022) uses a few chain-of-thought demonstrations as exemplars to improve the ability of LLMs to perform complex reasoning. Again, we employ a two-shot setting with one positive and one negative examples. (3) Chain-of-Logic (Servantez et al., 2024) elicits rule-based reasoning by decomposing the rule into elements, answering each rule element separately, and finally using a logical expression to obtain the final answer. This approach is meticulously designed for legal reasoning tasks.

All prompt template can be seen in Appendix C.

5.3 Experiment Results

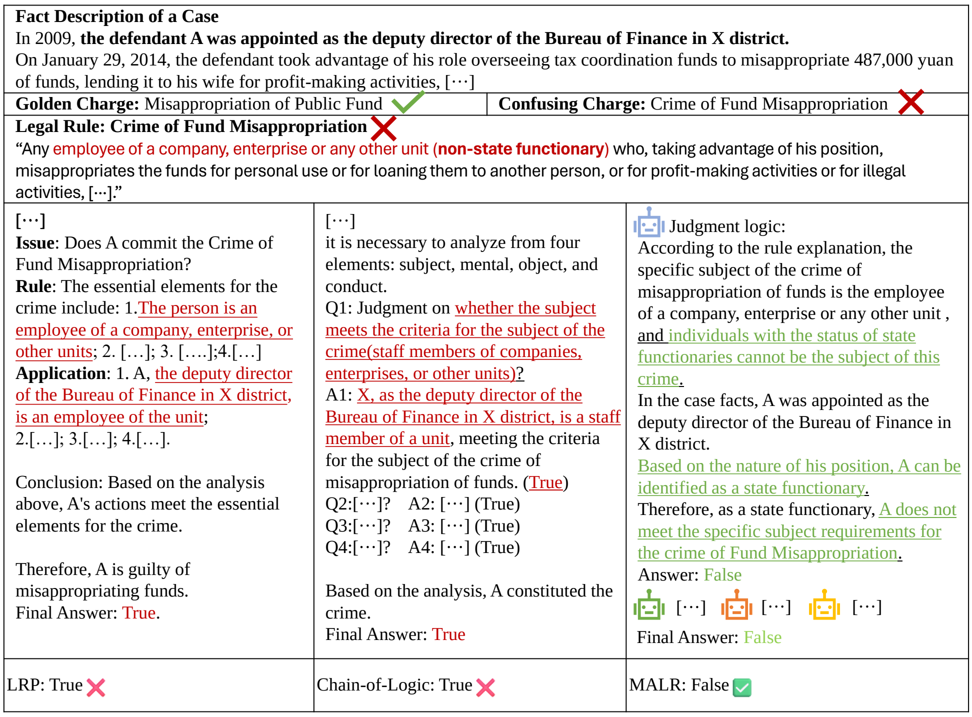
Main Results: From Table 1, we observe the following findings. (1) LLMs fail to distinguish confusing charges using simple but effective prompt methods such as CoT, and legal-specific prompting approaches also fail to predict accurately. By examining the actual prediction results, we found that LLMs using these methods tend to respond with "yes." (2) "MALR w/o insight", which only decomposes the task into sub-tasks, outperforms all the baselines. This result indicates that decomposing the task into sub-tasks may mitigate LLMs' biased tendencies. Notably, without any human intervention, our auto-planning strategy can decompose legal rules into four aspects: Subject (Sub), Mental (Men), Object (Obj) and Conduct (Con). This aligns with the Four Elements Theory (An et al., 2022), which is widely recognized in the legal domain. (3) "MALR w/o ask" does not utilize external knowledge feedback but still achieves the second-best results, indicating that the learned insights did significantly enhance the LLM's understanding of legal rules.(4) The complete MALR achieves the best performance on all datasets, demonstrating the effectiveness of proposed framework and the necessity of its core components. MALR achieves the best performance on nearly all confusing charge pairs (refer to Appendix D). (5) Regarding the base models, GPT-3.5 benefits more from our proposed MALR compared to GPT-4, achieving a more significant improvement over the baseline methods. This suggests that our framework has a stronger enhancing effect on LLMs with weaker foundational capabilities.

Ablation Results: Table 2 demonstrates the effectiveness of the components in adaptive rule-insights training module. (1) The results of "w/o $E_{success}$ (without Successful Experience)", "w/o E_{esp} (without Error-Success-Pair Experience)", and "w/o $M_{filtering}$ (without Insight Filtering)" prove the significance of each designed component in the learning from the trial-and-error process. (2) The "*directly generate*" approach involves encouraging the LLM to generate insights directly based on the legal rules without any training process. However, the performance drops in most situations, some-

[2] A legal-domain fine-tuned LLM based on Qwen(Bai et al., 2023), https://tongyi.aliyun.com/farui.

Methods	CAIL2018		CJO		CAIL-I	
	GPT-3.5	GPT-4	GPT-3.5	GPT-4	GPT-3.5	GPT-4
ZS-CoT (Kojima et al., 2022)	12.5	35.8	3.0	29.0	20.9	36.0
LRP (Yu et al., 2022)	9.8	37.8	1.0	37.0	22.3	49.6
FS-Prompt (Brown et al., 2020)	18.0	41.0	3.0	43.0	28.1	46.8
FS-CoT (Wei et al., 2022)	12.0	34.0	12.0	18.0	12.2	32.4
Chain-of-Logic (Servantez et al., 2024)	6.5	36.0	5.0	25.0	10.1	29.5
MALR w/o insight	32.3	43.8	22.0	44.0	45.3	53.2
MALR w/o ask	<u>37.3</u>	<u>53.3</u>	<u>31.0</u>	<u>53.0</u>	<u>51.1</u>	<u>55.4</u>
MALR (our)	**40.8**	**56.8**	**39.0**	**55.0**	**56.1**	**57.6**

Table 1: Main results on three legal datasets, the best is **bolded** and the second is <u>underlinded</u>. The metric is accuracy. w/o insight refers to only decompose to sub-tasks, w/o ask refers to do not get any external knowledge feedback.



Fact Description of a Case
In 2009, **the defendant A was appointed as the deputy director of the Bureau of Finance in X district.**
On January 29, 2014, the defendant took advantage of his role overseeing tax coordination funds to misappropriate 487,000 yuan of funds, lending it to his wife for profit-making activities, [···]

Golden Charge: Misappropriation of Public Fund ✓ | **Confusing Charge:** Crime of Fund Misappropriation ✗

Legal Rule: Crime of Fund Misappropriation ✗
"Any employee of a company, enterprise or any other unit (**non-state functionary**) who, taking advantage of his position, misappropriates the funds for personal use or for loaning them to another person, or for profit-making activities or for illegal activities, [···]."

[···]
Issue: Does A commit the Crime of Fund Misappropriation?
Rule: The essential elements for the crime include: 1.<u>The person is an employee of a company, enterprise, or other units</u>; 2. […]; 3. [….];4.[…]
Application: 1. A, <u>the deputy director of the Bureau of Finance in X district, is an employee of the unit</u>;
2.[…]; 3.[…]; 4.[…].

Conclusion: Based on the analysis above, A's actions meet the essential elements for the crime.

Therefore, A is guilty of misappropriating funds.
Final Answer: True.

[···]
it is necessary to analyze from four elements: subject, mental, object, and conduct.
Q1: Judgment on <u>whether the subject meets the criteria for the subject of the crime(staff members of companies, enterprises, or other units)?</u>
A1: <u>X, as the deputy director of the Bureau of Finance in X district, is a staff member of a unit</u>, meeting the criteria for the subject of the crime of misappropriation of funds. (<u>True</u>)
Q2:[···]? A2: [···] (True)
Q3:[···]? A3: [···] (True)
Q4:[···]? A4: [···] (True)

Based on the analysis, A constituted the crime.
Final Answer: True

Judgment logic:
According to the rule explanation, the specific subject of the crime of misappropriation of funds is the employee of a company, enterprise or any other unit , <u>and individuals with the status of state functionaries cannot be the subject of this crime.</u>
In the case facts, A was appointed as the deputy director of the Bureau of Finance in X district.
<u>Based on the nature of his position, A can be identified as a state functionary.</u>
Therefore, as a state functionary, <u>A does not meet the specific subject requirements for the crime of Fund Misappropriation.</u>
Answer: False
[···] [···] [···]
Final Answer: False

LRP: True ✗ | Chain-of-Logic: True ✗ | MALR: False ✓

Figure 4: Case study for a given case. The <u>green</u> parts mean are the most critical information for distinguish the confusing charges, the <u>red</u> parts are contents that do not match the facts of the case.

Datasets	CAIL2018		CJO	
Methods	GPT-3.5	**GPT-4**	GPT-3.5	**GPT-4**
w/o insights	32.3	43.8	22.0	44.0
w/o $E_{success}$	38.8	50.0	29.0	48.0
w/o E_{esp}	46.0	48.8	33.0	48.0
w/o $M_{filtering}$	38.0	54.0	31.0	53.0
directly generate	32.0	43.3	35.0	38.0
complete MALR	40.8	56.8	39.0	55.0

Table 2: Ablation test for adaptive rule-insights training module.

times even worse than without using insights at all. A possible explanation is that directly generating insights may lead to the inclusion of unimportant information. We provide case examples with explanations comparing the directly generated insights with those obtained through our training process in Appendix E.

Open-source LLMs with different model sizes. To further test the applicability of our MALR on different LMs (i.e. different sizes of open-source LLMs), we supplemented relevant experiments using a series of Qwen-2 models (Yang et al., 2024). Our findings indicate that our MALR achieves the best results across LLMs of different sizes and adheres to **scaling laws**. Interestingly, we also observed **more significant improvements in smaller LLMs**, which further demonstrates the effectiveness and practical significance of our proposed framework (details can be seen in Appendix F).

The Challenge and Significance of the Confus-

ing Charge Prediction Task: To demonstrate the challenge and significance of our proposed task, we thoroughly **compared General Charge Prediction and Confusing Charge Prediction**. These comparisons clearly indicate that while general legal models perform well on traditional general charge prediction tasks, they are less effective for the confusing charge prediction task. Additionally, we also **analyzed the performance of human annotators** on this task. Our findings demonstrate the urgency and importance of this task setting, and they reveal that MALR can even surpass human performance (details can be seen in Appendix G).

5.4 Case Study

Figure 4 presents an example of different methods used to predict confusing charges. As demonstrated in the case, our framework effectively focuses on the most critical aspects of the legal rules and makes a well-reasoned judgment. In contrast, both LRP and Chain-of-Logic overlook the crucial information in the legal rules, resulting in their failure to accurately predict the confusing charge.

6 Conclusion

In the study, we introduce a challenging task to better evaluate LLMs' capability to comprehend legal theories. The proposed MALR framework can automatically decomposes complex legal tasks and extracts insights from legal rules, enhancing LLMs' legal reasoning abilities. Extensive experiments demonstrate MALR's effectiveness in equipping LLMs with a robust understanding of legal rules.

7 Ethical Considerations

The datasets we used for evaluation are all from public legal datasets, and information about the defendants has been anonymized. To ensure personal privacy is not violated, we conduct a secondary review before releasing our dataset to ensure all personal information has been completely removed.

Our work focuses on exploring algorithms to enhance the complex reasoning capabilities of LLMs, rather than replacing human judges or being directly used in real-world decision-making applications. In practical use, human judges should act as the final safeguard to maintain fairness and mitigate the potential harms related to algorithms. We will restrict its use for non-commercial purposes such as academic research through a specific license.

8 Limitations

Our work has two main limitations. First, even though we achieved great results, MALR did not predict correctly on all confusing charge pair cases. In the future, retrieval augmented generation could help our model perform better.

Second, our framework shows that LLMs can self-improve by summarize insights into the rules from trials and errors, which helps LLMs to better perform in complex legal reasoning tasks. Nevertheless, the potential for applying this approach in other fields such as medicine, finance, and scientific discovery remains unexplored. In the future, our framework could be applied in diverse domains.

Acknowledgments

This work is supported by Alibaba Innovative Research Program. This work is supported by the National Natural Science Foundation of China (72104212, 62106039), the Natural Science Foundation of Zhejiang Province (LY22G030002), and the Fundamental Research Funds for the Central Universities.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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, Florence, Italy. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Simucourt: Building judicial decision-making agents with real-world judgement documents. *arXiv preprint arXiv:2403.02959*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press.

Ang Li, Qiangchao Chen, Yiquan Wu, Xiang Zhou, Kun Kuang, Fei Wu, and Ming Cai. 2024. From graph to word bag: Introducing domain knowledge to confusing charge prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7469–7479, Torino, Italia. ELRA and ICCL.

Dugang Liu, Weihao Du, Lei Li, Weike Pan, and Zhong Ming. 2022. Augmenting legal judgment prediction with contrastive case relations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2658–2667.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

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

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When do llms need retrieval augmentation? mitigating llms' overconfidence helps retrieval augmentation. *arXiv preprint arXiv:2402.11457*.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. 2023. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024. Chain of logic: Rule-based reasoning with large language models. *arXiv preprint arXiv:2402.10400*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Changlong Sun, Yating Zhang, Xiaozhong Liu, and Fei Wu. 2020. Legal intelligence: Algorithmic, data, and social challenges. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2464–2467, New York, NY, USA. Association for Computing Machinery.

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, Torino, Italia. ELRA and ICCL.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

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

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2024. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024. Can llms reason with rules? logic scaffolding for stress-testing and improving llms. *arXiv preprint arXiv:2402.11442*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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, Singapore. Association for Computational Linguistics.

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*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 973–982.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3540–3549.

A Prompt Template for our MALR Agents

The prompt Templates for each agents can refer to Figure 5, 6, 7. We provide prompt templates in English; however, when applied in practice, these templates can be adapted to different languages by translating them into the corresponding language-specific prompts.

B Dataset and Experiments Information

CAIL2018 is a popular Chinese charge prediction datasets. It consists of real-world cases, each of which includes a fact description and the corresponding charges labels.

CJO is another Chinese legal dataset, same source from the CAIL2018, which is constructed to mitigate the potential data leakage.

CAIL-I contains 462 innocent cases that did not involve any crime. The dataset also has annotations for the criminal charge most similar to the non-criminal facts. The legal judgment prediction for an innocent case adheres to the presumption of innocence. It can evaluate whether LLMs can fully conform to legal rules for reasoning.

Key differences between each pair of confusing charges are provided in Figure 8. **Model Cost**. Statistically, the total token of our method is 1365 for each CAIL2018 example and the inference time per example is about 22s.

C Prompt Template for Baseline

The prompt templates for each baseline can refer to Figure 9, 10, 11. We provide prompt templates in English; however, when applied in practice, these templates can be adapted to different languages by translating them into the corresponding language-specific prompts.

Methods	Qwen-2-1.5B	Qwen-2-7B	Qwen-2-57B-A14B	Qwen-2-72B
ZS-CoT	16.3	6.25	16	21
LRP	14.8	6	13	26.3
FS-Prompt	7.8	29	35	48.5
FS-CoT	10.8	12.5	20.5	33
Chain-of-Logic	2.3	8.3	15.8	28
MALR w/o insight	21.8	21	22.8	40
MALR w/o ask	24	33.3	31.5	48.8
MALR (our)	**27.3**	**38.3**	**40**	**50.5**
Improvement	67.7%	31.9%	14.3%	4.1%

Table 3: Performance of Qwen-2 models with different sizes on CAIL2018 dataset. The "Improvement" shows the performance improvement of our MALR compared to the strongest baseline.

D Specific Performance in the CAIL2018 dataset

Table 4 details the specific performance for each confusing-charge pair in the CAIL2018 dataset. The proposed MALR framework achieves the best performance on nearly all confusing charge pairs.

E More Cases for our insights

Figure 12 shows our training rule-insights can better learn the slight difference in the legal rules, which encourage the LLMs to better understand the legal rules.

F MALR performance in Open-source LLMs

We applied Qwen-2 (Yang et al., 2024) in our framework for additional test. Qwen-2 includes a series of models with different sizes. We selected models with sizes of Qwen-2-1.5B, Qwen-2-7B, Qwen-2-57B-A14B, and Qwen-2-72B. The results on the CAIL-2018 dataset are as shows in Table 3:

We also observed some interesting phenomena:

Scaling Law: For models of different sizes, performance gradually improves as the model size increases. This is consistent with the conclusions of the scaling law and proves the robustness of our proposed model, which works effectively across models of varying sizes.

Significant Improvements for Smaller LLMs: The MALR method may offer more substantial improvements for LLMs with relatively weaker foundational capabilities. The smallest model, Qwen-2-1.5B, has a 67.7% improvement over the strongest baseline with our MALR. In contrast, the 72B Qwen model only showed a 4.1% improvement over the highest baseline. This indicates that our framework can bring significant enhancements when the LLM's capacity is relatively limited. This is crucial for real-world applications, as not all institutions and individuals can afford the most powerful LLMs.

G Challenge and Significance of our proposed task

We conducted extensive experiments and human evaluations. The results demonstrate the urgency and importance of our Confusing Charge Prediction task setting and show that MALR can even outperform human performance in our task.

Comparisons to General Charge Prediction. Confusing Charge Prediction (An et al., 2022) is motivated by the challenge that legal prediction models encounter when distinguishing between charges with similar meanings in real legal scenarios. Compared to General Charge Prediction, Confusing Charge Prediction (Xiao et al., 2018) is significant and valuable because it tackles a critical and practical challenge. In real legal scenarios, legal models often confuse and misinterpret charges with subtle differences. Enhancing the ability of models to accurately predict these confusing charges can improve the accuracy and reliability of legal AI systems.

From the perspective of task formulation, the General Charge Prediction Task is a multi-label classification problem where the input is a fact description and the output is one of the charge categories. Our Confusing Prediction Task can be described as follows: For a given fact description, the prediction model aims to determine if it satisfies the rule of the golden charge while not matching the rule of a confusing charge.

First, based on the typical task setup for General Charge Prediction, we performed a multi-label classification comparison. We trained Lawformer (Xiao et al., 2021) on the entire CAIL-2018 training set. The model achieved an overall accuracy of 85.15% on the test set. In contrast, the average prediction accuracy for the 8 pairs of confusing charges selected in our paper was only 72.19% (nearly 13% lower). For example, for the charge of "Misappropriation of Funds" the model's accuracy was 76.45%. Among the misclassified samples, 51.35% were predicted as a confusing charge "Misappropriation of Public Funds". Similarly, for the charge of "Bribery by Non-State Officials", the model's accuracy was 70.54%. Among the misclassified samples, 58.84% were predicted as a confusing charge of "Bribery". In contrast, certain charges that are relatively easy to distinguish, such as "Illegal Cultivation of Drug Plants" had an accuracy of 99.95%, and "Environmental Pollution" had an accuracy of 99.3%.

Second, We evaluated two small legal language models (legal-xlm-roberta-base (Niklaus et al., 2023) and lawformer(Xiao et al., 2021)) based on the CAIL-2018 dataset, maintaining consistent confusing charge prediction task settings. The evaluation results were as follows: the accuracy of Lawformer was 2%, and the accuracy of legal-xlm-roberta-base was 2.25%. In contrast, our proposed MALR leverages the generalization and comprehension abilities of LLMs, highlighting the advantages of our framework.

Comparisons to Human Annotation. To further validate the quality of our proposed task and the effectiveness of the MALR framework, we compared the results of the LLM with those achievable by humans. We extracted 20 case facts from the CAIL-2018 dataset and randomly selected either the golden charge or a confusing charge, along with the corresponding legal rules for each case. Our human evaluation follows standard LLM assessment practices: ensuring annotators have comprehension and reasoning skills, with minimal prior knowledge of the answers. Therefore, we recruited 6 annotators, all with bachelor's degrees and no legal background. They will receive the necessary training to complete the tasks, and during the evaluation process, they will be provided with legal rules to aid their reasoning, ensuring consistency with the LLM's input.

The human annotators achieved an average accuracy of 62.5%, with an average completion time of 28.3 minutes. The LLM baseline methods achieved an average accuracy of 53%, with an average completion time of 8.2 minutes. The accuracy of our proposed MALR is 65%, with a total execution time of 10.6 minutes. Our proposed MALR framework not only surpasses the average performance of human annotators in terms of accuracy but also demonstrates superior efficiency in execution time. These results further prove the effectiveness and efficiency of our methodology.

Auto-Planner

You are currently in the task planning stage. Given a [Legal Rule Description] and related [Fact Descriptions of the case]. Please break it down into sub-tasks.
[Legal Rule Description]
{legal rule}

[Fact Descriptions of the case]
{fact description}

- Each sub-task action MUST have a unique ID, which is strictly increasing.
- Ensure the plan maximizes parallelizability.
- Never explain the sub-task actions with comments.

Sub-task Agent

You are a helpful legal profession. With a clear definition of the rule of {sub-task}.
Please determine whether {criminals} commit the crime of {charge_name} on the {sub-task} aspect, based on the [Legal Rule Description] and [Fact Descriptions of the case].
(Answer True if it constitutes the crime of {charge_name}, and answer False if it does not).

[Legal Rule Description]
{legal rule on sub-task}
Note: {rule-insights into the sub-task legal rule} //When training insights, this is Empty String//

[Fact Descriptions of the case]
{fact description}
[Knowledge feedback based on insight]
{Knowledge_feedback_by_external expert} //When training insights, this is Empty String//
Note:
Clarify the elements of {sub-task} and their corresponding relationship with the rules, clearly express your judgment logic, and provide a definite conclusion answer: True, False (answer True if it constitutes the {sub-task} of the crime of {charge_name}, answer False if it does not).
Answer format: [Judgment Logic] + [Answer]

Self-Reflector

You are an advanced legal agent who can analyze the incorrect answer and reasons through self-reflection.
By breaking down the task into following sub-tasks: {sub-task list}, sub-task experts reason that whether the defendant commits the crime of certain charge on the corresponding sub-aspect, based on the [Sub-task Legal Rule] and [Fact Descriptions of the case].
But sub-task experts incorrectly answered the question, please analyze where the judgment was mistatken based on the error trial, which could be one or more sub-tasks.

[Legal Rule Description]
{legal rule on sub-task}

[Fact Descriptions of the case]
{fact description}

[Incorrect Answer]
{initial_error_answers}

[ground truth]
{GROUND TRUTH FROM EXERTNAL FEEDBACK}

[Requirement]
[answer format]
Aspect1: <ONLY the option word of the four aspects; not a complete sentence!>
Reason1: <ONLY the reason why Aspect1 you conclude error results in Chinese>
...
Select the key error aspect, NOT all aspects are necessary to analyze.

Figure 5: Prompt Template for Auto-Planer, Sub-task Agent and Self-Reflector

Insight Drawer for error-success-pair experience

You are an advanced legal agent who can draw insight into the rule to improve by self-reflection.
I will give your two attempts at answering a legal reasoning question based on a given the [Legal Rule Description] and [Fact Descriptions of the case].
There are one incorrect answer and one correct answer. Please generate one-sentence insight into the sub-task legal rule to highlight the most critical task-level judgment factor, NOT mention any specific information like defendant's name.

[Legal Rule Description]
{legal rule on error sub-task}

[Fact Descriptions of the case]
{fact description}

[Question]
Please determine whether {criminal} commit the crime of {charge_name} on the {sub-task} aspect, based on the [Legal Rule Description] and [Fact Descriptions of the case].

[Error Trial]
{error_trial}

[Success Trial]
{success_trial}

[Output]

Insight Drawer for successful experience

You are an advanced legal agent who can draw insight into the rule to improve by self-reflection.
I will give your two attempts regarding the judgment of a case. The first is to answer whether the fact meets [Legal Rule of {golden_charge}], and the second is whether it meets [Legal Rule of {confuing_charge}].

Please generate one-sentence insight into the rule to highlight the most critical task-level judgment factor between the two charges. NOT mention any specific information like defendant's name.

[Fact Descriptions of the case]
{fact description}

[Legal Rule Description 1]
{Golden charge's legal rule}

[Question]
Please determine whether {criminal} commit the crime of {golden_charge}
[Answer]
{Successful Trial for all sub-tasks responses}

[Legal Rule Description 2]
{Confusing charge's legal rule}

[Question]
Please determine whether {criminal} commit the crime of {confusing_charge}
[Answer]
{Successful Trial for all sub-tasks responses}

[Output]

Figure 6: Prompt Template for Insight Drawer

Insight Filtering

You are an insight filtering who can filter the insights in the rule-insight knowledge base.

[Insights knowledge base]
{insight_from_knowledge_base} /JSON Format/

[Requirement]

1. Check the correctness for insights
2. Filter and remove duplicate insights
3. Don’t change the original expression of any insights
4. Return the same json format as [Insights knowledge base]

Insight Inferencer

You are an expert at extracting the most critical information from rules, and you will be given some legal rules and the insights that have been extracted from them.
These insights can help judges make court decisions.
Please refer to the following rules and insights, and generate the corresponding insight within a new legal rule.

[Example 1]
Legal Rule:
{similar_rule}

Insight:
{similar_rule_insight}

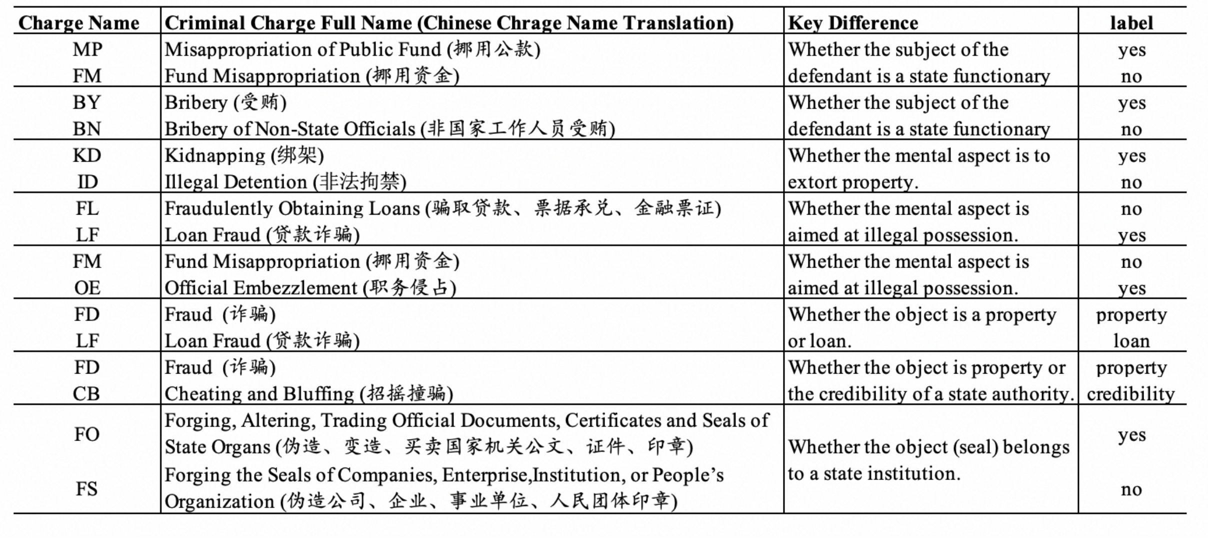
[Your turn]
Legal Rule:
{new_charge_rule}
Insight:

Ask Key Question for Fact Checking

Please form a key question based on the [insight] and [case fact].
[Start of Examples]
[insight]
If the subject is a state functionary, it does not meet the subject criteria for the crime of fund misappropriation.
[case fact]
[…]The defendant, taking advantage of his position as a customer manager at the X of the Agricultural Bank of China XXX, misappropriated RMB 400,000 of the unit's funds under the name of loan customer XXX by forging materials required for the "second use of credit application" of a business loan in XXX name on January 6, 2015.
[…]
[Question]
Does the defendant qualify as the subject for the crime of fund misappropriation?
[Your response]
S1: Review of the subject for the crime of fund misappropriation: The defendant is a customer manager at X of the Agricultural Bank of China XXX.
S2: Relationship between the subject and the insight: If the subject is a state functionary, it does not meet the subject criteria for the crime of fund misappropriation.
S3: Therefore, the key question formed is: Is the customer manager at X of the Agricultural Bank of China XXX a state functionary?
[End of Examples]

[Your turn]
[insight] {insight}
[case fact] {fact}
[question] Does this case constitute the element of {charge_name}?
[Your response]

Figure 7: Prompt Template for Insight Filtering, Insight Inference and Ask Key Question for Fact Checking

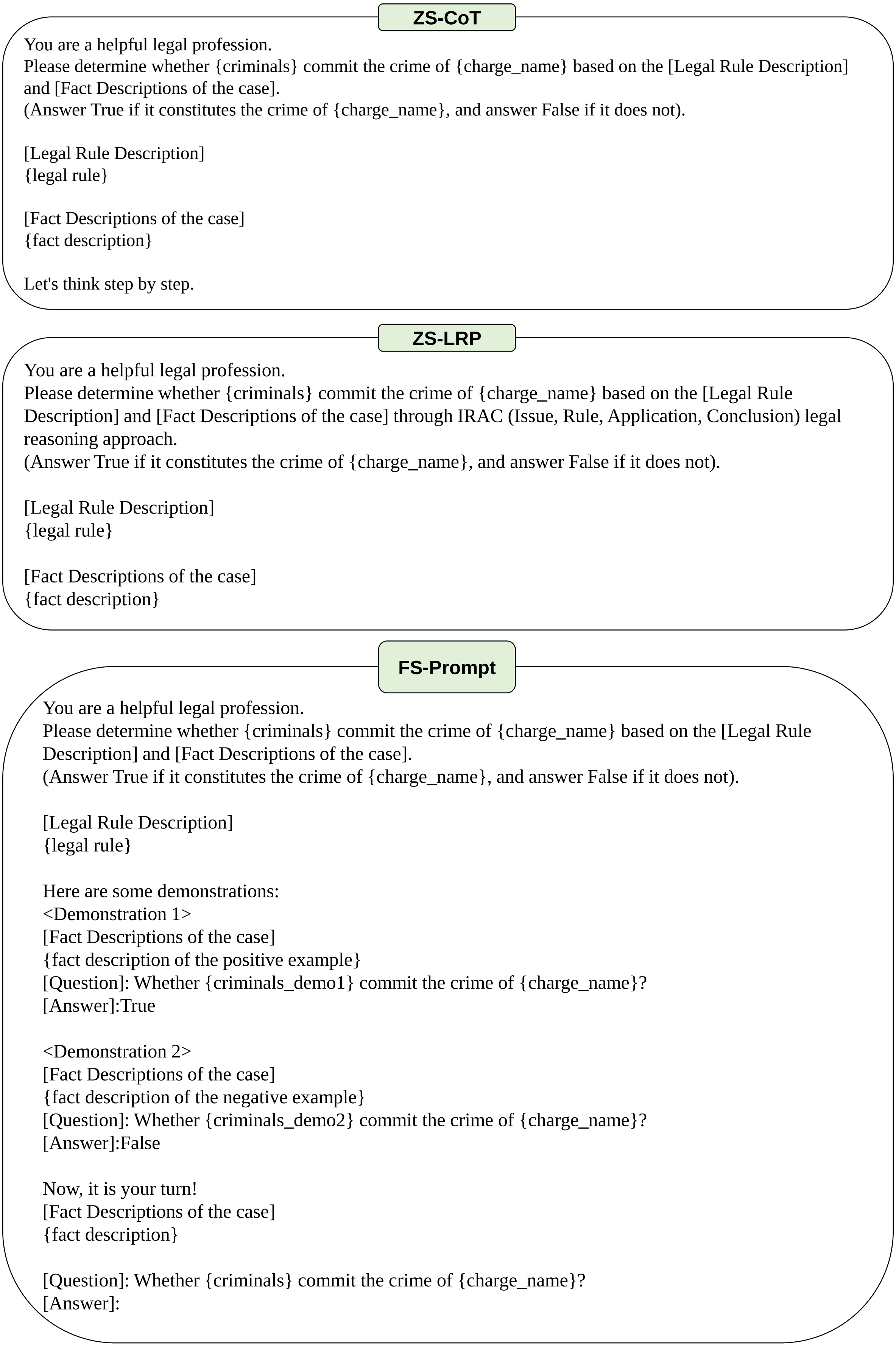


Charge Name	Criminal Charge Full Name (Chinese Chrage Name Translation)	Key Difference	label
MP	Misappropriation of Public Fund (挪用公款)	Whether the subject of the	yes
FM	Fund Misappropriation (挪用资金)	defendant is a state functionary	no
BY	Bribery (受贿)	Whether the subject of the	yes
BN	Bribery of Non-State Officials (非国家工作人员受贿)	defendant is a state functionary	no
KD	Kidnapping (绑架)	Whether the mental aspect is to	yes
ID	Illegal Detention (非法拘禁)	extort property.	no
FL	Fraudulently Obtaining Loans (骗取贷款、票据承兑、金融票证)	Whether the mental aspect is	no
LF	Loan Fraud (贷款诈骗)	aimed at illegal possession.	yes
FM	Fund Misappropriation (挪用资金)	Whether the mental aspect is	no
OE	Official Embezzlement (职务侵占)	aimed at illegal possession.	yes
FD	Fraud (诈骗)	Whether the object is a property	property
LF	Loan Fraud (贷款诈骗)	or loan.	loan
FD	Fraud (诈骗)	Whether the object is property or	property
CB	Cheating and Bluffing (招摇撞骗)	the credibility of a state authority.	credibility
FO	Forging, Altering, Trading Official Documents, Certificates and Seals of State Organs (伪造、变造、买卖国家机关公文、证件、印章)	Whether the object (seal) belongs to a state institution.	yes
FS	Forging the Seals of Companies, Enterprise,Institution, or People's Organization (伪造公司、企业、事业单位、人民团体印章)		no

Figure 8: Key difference between each pair of confusing charge

Golden Charge	MP	FM	BY	BN	KD	ID	FL	LF	FM	OE	FD	LF	FD	CB	FO	FS
GPT-3.5																
ZS-CoT	4.0	0.0	12.0	8.0	32.0	4.0	0.0	0.0	0.0	16.0	36.0	0.0	24.0	8.0	56.0	0.0
LRP	0.0	0.0	0.0	0.0	32.0	0.0	4.0	0.0	0.0	16.0	20.0	4.0	20.0	4.0	48.0	8.0
FS-Prompt	0.0	0.0	0.0	0.0	40.0	8.0	0.0	0.0	**8.0**	8.0	**76.0**	4.0	**72.0**	4.0	**68.0**	0.0
FS-CoT	8.0	**64.0**	12.0	0.0	20.0	0.0	0.0	0.0	0.0	12.0	36.0	0.0	16.0	0.0	24.0	0.0
Chain-of-Logic	0.0	28.0	0.0	8.0	0.0	8.0	0.0	0.0	4.0	0.0	4.0	0.0	28.0	0.0	20.0	40.0
MALR (Our)	**24.0**	**64.0**	**64.0**	**16.0**	**68.0**	**28.0**	**28.0**	**12.0**	**8.0**	**28.0**	24.0	**72.0**	32.0	**44.0**	52.0	**88.0**
GPT-4																
ZS-CoT	12.0	52.0	68.0	12.0	24.0	0.0	0.0	0.0	4.0	**40.0**	96.0	4.0	96.0	8.0	76.0	80.0
LRP	20.0	76.0	60.0	**32.0**	16.0	44.0	8.0	0.0	28.0	24.0	80.0	8.0	80.0	16.0	56.0	60.0
FS-Prompt	12.0	56.0	**84.0**	**32.0**	20.0	0.0	16.0	**60.0**	**88.0**	20.0	56.0	32.0	92.0	**20.0**	40.0	28.0
FS-CoT	8.0	64.0	48.0	12.0	24.0	0.0	0.0	0.0	0.0	20.0	**100.0**	0.0	92.0	4.0	**84.0**	88.0
Chain-of-Logic	8.0	80.0	56.0	16.0	24.0	0.0	0.0	0.0	8.0	16.0	**100.0**	0.0	92.0	12.0	80.0	84.0
MALR (Our)	**36.0**	**88.0**	**84.0**	**32.0**	**36.0**	**76.0**	**32.0**	28.0	44.0	20.0	96.0	**56.0**	**100.0**	12.0	72.0	**96.0**

Table 4: Results on each criminal charge of confusing-charge pairs on CAIL2018 dataset.



ZS-CoT

You are a helpful legal profession.
Please determine whether {criminals} commit the crime of {charge_name} based on the [Legal Rule Description] and [Fact Descriptions of the case].
(Answer True if it constitutes the crime of {charge_name}, and answer False if it does not).

[Legal Rule Description]
{legal rule}

[Fact Descriptions of the case]
{fact description}

Let's think step by step.

ZS-LRP

You are a helpful legal profession.
Please determine whether {criminals} commit the crime of {charge_name} based on the [Legal Rule Description] and [Fact Descriptions of the case] through IRAC (Issue, Rule, Application, Conclusion) legal reasoning approach.
(Answer True if it constitutes the crime of {charge_name}, and answer False if it does not).

[Legal Rule Description]
{legal rule}

[Fact Descriptions of the case]
{fact description}

FS-Prompt

You are a helpful legal profession.
Please determine whether {criminals} commit the crime of {charge_name} based on the [Legal Rule Description] and [Fact Descriptions of the case].
(Answer True if it constitutes the crime of {charge_name}, and answer False if it does not).

[Legal Rule Description]
{legal rule}

Here are some demonstrations:
<Demonstration 1>
[Fact Descriptions of the case]
{fact description of the positive example}
[Question]: Whether {criminals_demo1} commit the crime of {charge_name}?
[Answer]:True

<Demonstration 2>
[Fact Descriptions of the case]
{fact description of the negative example}
[Question]: Whether {criminals_demo2} commit the crime of {charge_name}?
[Answer]:False

Now, it is your turn!
[Fact Descriptions of the case]
{fact description}

[Question]: Whether {criminals} commit the crime of {charge_name}?
[Answer]:

Figure 9: Prompt Template for baseline ZS-CoT, ZS-LRP and FS-Prompt

FS-CoT

You are a helpful legal profession.
Please determine whether {criminals} commit the crime of {charge_name} based on the [Legal Rule Description] and [Fact Descriptions of the case].
(Answer True if it constitutes the crime of {charge_name}, and answer False if it does not).

Here are some demonstrations:
<Demonstration 1>
[Legal Rule Description]
{legal rule}
[Fact Descriptions of the case]
{fact description of the positive example}
[Question]: Whether {criminals_demo1} commit the crime of {charge_name}?
[Judgment Logic]:
{chain_of_thought_for_demo1}
[Answer]: True

<Demonstration 2>
[Legal Rule Description]
{legal rule}
[Fact Descriptions of the case]
{fact description of the negative example}
[Question]: Whether {criminals_demo2} commit the crime of {charge_name}?
[Judgment Logic]:
{chain_of_thought_for_demo2}
[Answer]: False

Now, it is your turn!
[Legal Rule Description]
{legal rule}

[Fact Descriptions of the case]
{fact description}

[Question]: Whether {criminals} commit the crime of {charge_name}?
[Judgment Logic]:

Figure 10: Prompt Template for baseline FS-CoT

Chain-of-Logic

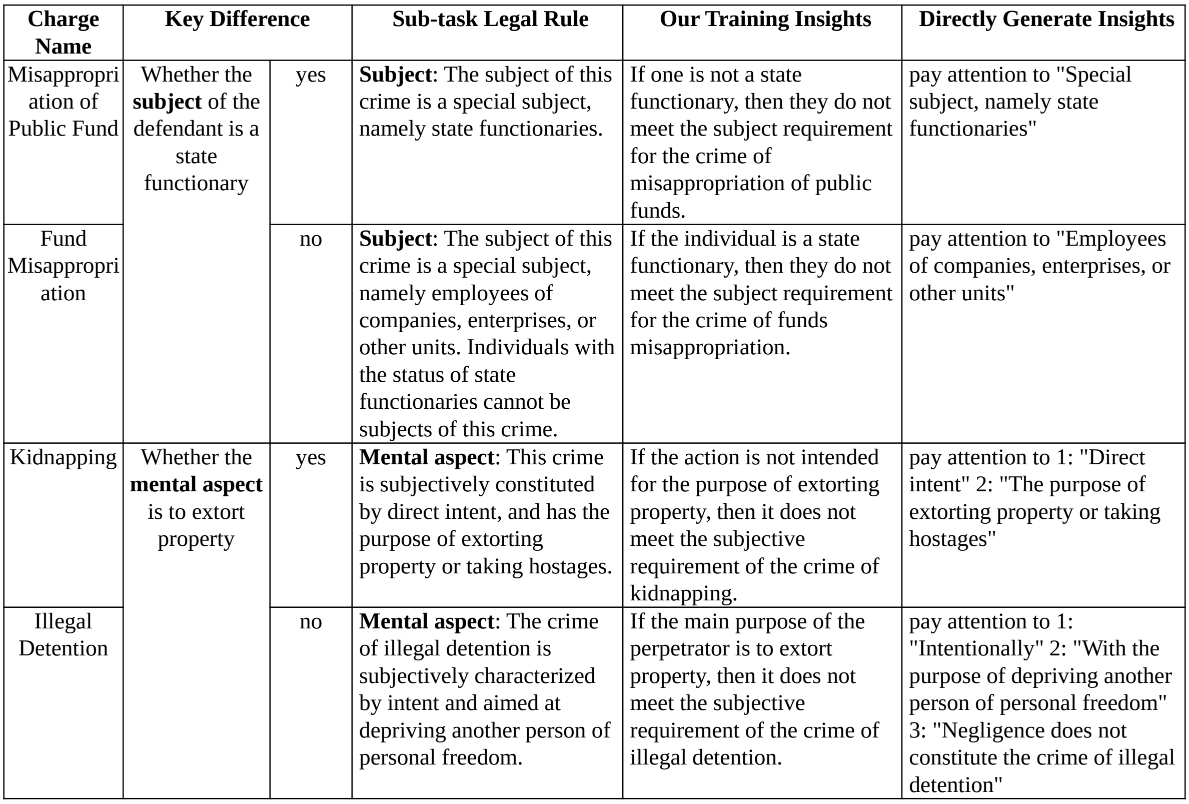
You are a helpful legal profession.
Please determine whether {criminals} commit the crime of {charge_name} based on the [Legal Rule Description] and [Fact Descriptions of the case].
(Answer True if it constitutes the crime of {charge_name}, and answer False if it does not).

Here are some demonstrations:
<Demonstration 1>
[Legal Rule Description]
{legal rule}
[Fact Descriptions of the case]
{fact description of the positive example}
[Question]: Whether {criminals_demo1} commit the crime of {charge_name}?
[Judgment Logic]:
Decompose the rule into elements:
The rule can be decomposed by (A) subject rule, (B) mental aspect rule, (C) object rule, (D) conduct aspect rule.
Logical Expression: (A and B and C and D)
Answer each rule element separately:
Q1: Does the defendant satisfy the subject rule (specific content in the subject rule of {charge_name})?
A1:The defendant is the xx, so satisfied the subject rule.(True)
Q2: Does the defendant satisfy the mental aspect rule (specific content in the mental aspect rule of {charge_name})?
A2:The defendant is the xx, so satisfied the mental aspect rule.(True)
Q3: Does the defendant satisfy the object rule (specific content in the object rule of {charge_name})?
A3:The defendant is the xx, so satisfied the object rule.(True)
Q4: Does the defendant satisfy the conduct aspect rule (specific content in the conduct aspect rule of {charge_name})?
A4:The defendant is the xx, so satisfied the conduct aspect rule.(True)
Logical expression with answer: (True and True and True and True) = True
So the defendant commits the crime of {charge_name}.
[Answer]The final answer is: True

<Demonstration 2>
[Legal Rule Description]
{legal rule}
[Fact Descriptions of the case]
{fact description of the positive example}
[Question]: Whether {criminals_demo2} commit the crime of {charge_name}?
[Judgment Logic]:
…
//The Judgment Logic format is similar to the Demonstration 1//
…
Logical expression with answer: (False and True and True and True) = False
So the defendant does not commits the crime of {charge_name}.
[Answer]The final answer is: False

Now, it is your turn!
[Legal Rule Description]
{legal rule}
[Fact Descriptions of the case]
{fact description}
[Question]: Whether {criminals} commit the crime of {charge_name}?
[Judgment Logic]:

Figure 11: Prompt Template for baseline Chain-of-Logic



Charge Name	Key Difference		Sub-task Legal Rule	Our Training Insights	Directly Generate Insights
Misappropriation of Public Fund	Whether the **subject** of the defendant is a state functionary	yes	**Subject**: The subject of this crime is a special subject, namely state functionaries.	If one is not a state functionary, then they do not meet the subject requirement for the crime of misappropriation of public funds.	pay attention to "Special subject, namely state functionaries"
Fund Misappropriation		no	**Subject**: The subject of this crime is a special subject, namely employees of companies, enterprises, or other units. Individuals with the status of state functionaries cannot be subjects of this crime.	If the individual is a state functionary, then they do not meet the subject requirement for the crime of funds misappropriation.	pay attention to "Employees of companies, enterprises, or other units"
Kidnapping	Whether the **mental aspect** is to extort property	yes	**Mental aspect**: This crime is subjectively constituted by direct intent, and has the purpose of extorting property or taking hostages.	If the action is not intended for the purpose of extorting property, then it does not meet the subjective requirement of the crime of kidnapping.	pay attention to 1: "Direct intent" 2: "The purpose of extorting property or taking hostages"
Illegal Detention		no	**Mental aspect**: The crime of illegal detention is subjectively characterized by intent and aimed at depriving another person of personal freedom.	If the main purpose of the perpetrator is to extort property, then it does not meet the subjective requirement of the crime of illegal detention.	pay attention to 1: "Intentionally" 2: "With the purpose of depriving another person of personal freedom" 3: "Negligence does not constitute the crime of illegal detention"

Figure 12: Case study for illustrating the effectiveness of our training insights.