Legal Rule Induction: Towards Generalizable Principle Discovery from Analogous Judicial Precedents

Anonymous ACL submission

Abstract

Legal rules encompass not only codified statutes but also implicit adjudicatory principles derived from precedents that contain discretionary norms, social morality, and policy. While computational legal research has advanced in applying established rules to cases, *inducing* legal rules from judicial decisions remains understudied, constrained by limitations in model inference efficacy and symbolic reasoning capability. The advent of Large Language Models (LLMs) offers unprecedented opportunities for automating the extraction of such latent principles, yet progress is stymied by the absence of formal task definitions, benchmark datasets, and methodologies. To address this gap, we formalize Legal Rule Induction (LRI) as the task of deriving concise, generalizable doctrinal rules from sets of analogous precedents, distilling their shared preconditions, normative behaviors, and legal consequences. We introduce the first LRI benchmark, comprising 5,121 case sets (38,088 Chinese court cases in total) for model tuning and 216 expert-annotated gold test sets. Experimental results reveal that: 1) State-of-the-art LLMs struggle with over-generalization and hallucination; 2) Training on our dataset markedly enhances LLMs' capabilities in capturing nuanced rule patterns across similar cases.

"Common law courts have two functions: resolving disputes according to legal rules and making legal rules."

— Melvin A. Eisenberg

1 Introduction

Modern legal systems, whether grounded in statutory codes or the case-law tradition, ultimately reason through legal rules (Eisenberg, 2022). In civil law jurisdictions (*e.g.*, China and France) (Merryman and Pérez-Perdomo, 2018; Watkin, 2017), rules are codified in statutory provisions characterized by explicit logical structures (Lei, 2013).

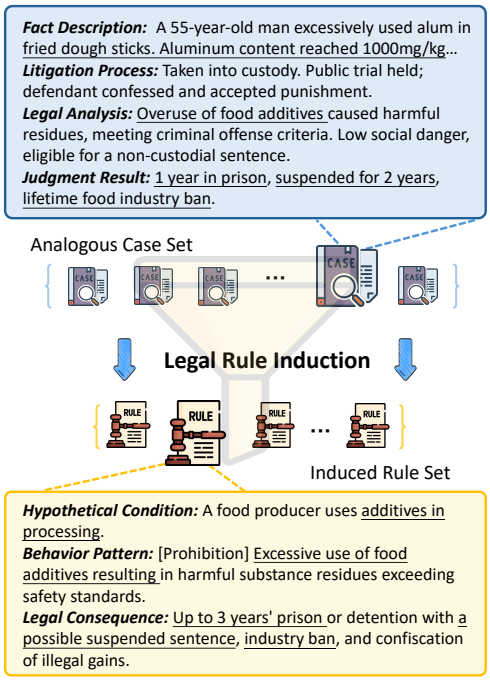

Figure 1: An illustration of legal rule induction from analogous judicial cases via the three-element logical structure of legal rules (Wenxian et al., 2018).

Common law systems, while also referencing statutes, primarily operationalize rules through precedent (Holmes Jr, 2020). Under stare decisis (Douglas, 1949), a court is obliged to apply the rule articulated in any binding precedent, whether issued by a higher court or by itself, whenever the present case is materially indistinguishable (Eisenberg, 2022). Although these systems differ superficially, explicit code articles versus implicit precedent rule (Brewer, 2013; Lamond, 2005), civil and common law rely on the same normative atom: the legal rule (Dickinson, 1931). Hence, the capacity to extract, articulate, and employ that atom is indispensable to any form of legal reasoning (Levi, 2013; Guha et al., 2023).

Current computational legal research tends to bifurcate statutory and precedent-based reasoning, of-

ten framing the former as primarily **deductive reasoning** (Blair-Stanek et al., 2023) (applying statutory rules to specific facts) and the latter as relying on similarity matching (Liu and Zheng, 2025), neglecting their common grounding in rules. This leaves legal **inductive reasoning** (*i.e.*, rule induction), the vital link between these approaches and a cornerstone of everyday legal work, critically underexplored. As Melvin Eisenberg highlights, a common law court performs two critical functions: resolving disputes by applying established rules and, crucially, formulating new rules from clusters of earlier decisions (Eisenberg, 2022). Additionally, lawyers, pro se litigants, and judges spend considerable effort sifting through massive corpora of opinions or judgments to extract abstract propositions that support their positions. The advent of Large Language Models (LLMs) (DeepSeek-AI et al., 2025b; Achiam et al., 2024; Qwen et al., 2025), with their extensive context windows and impressive reasoning capabilities, raises the possibility of automatic rule induction from lengthy judicial documents. Yet the task remains underdefined and essentially unsolved: there is no precise task definition, no public dataset, and no standard methodology.

To bridge this gap, we formally propose the **L**egal **R**ule **I**nduction **(LRI)** task, defined as the synthesis of abstract legal rules from analogous judicial precedents, as illustrated in Figure 1. Informed by jurisprudence in China (Wenxian et al., 2018), we define a legal rule by three core elements: hypothetical applicability conditions triggering the rule, behavioral prescriptions that govern conduct (permitting, prohibiting, or obligating actions), and legal consequences specifying outcomes, whether positive (*e.g.*, rights conferred) or negative (*e.g.*, punishments imposed). Input precedents for the LRI task consist of facts, procedural history, legal analysis, and judgment, excluding statutory citations to compel models towards genuine rule induction rather than mere recall of codified law (Louis et al., 2023a).

To facilitate LRI research and benchmark LLM performance, we introduce the **LRI Dataset**, a large-scale corpus specifically constructed for rule induction studies. In common law systems, legal rules in judgments are derived both from explicit statutory citations and from legal reasoning embedded in precedents. While statutory rules are relatively accessible, extracting implicit rules from precedents requires labor-intensive expert effort. In contrast, civil law judgments typically cite and apply codified statutes directly, enabling more straightforward case-to-rule alignment. Exploiting this feature, we focus on *China Judgments Online*[1], one of the largest and most authoritative data sources in civil law jurisdictions, from which we collect over 9 million original civil and criminal cases and cluster them into case sets that reference the same statutory articles. Each resulting set thereby shares explicit grounding in statutory rules while also revealing, through the courts' analyses, any implicit discretionary principles applied. Following an automated processing pipeline via DeepSeek-R1 (DeepSeek-AI et al., 2025a) and applying filters based on set size and rule applicability, we curate the **LRI-AUTO** dataset of 5,121 case sets (comprising 38,088 judgments) for model tuning.

For rigorous evaluation, we further develop **LRI-GOLD**, a meticulously curated test set composed of 216 case sets (1,620 cases) annotated by legal researchers. Our experimental evaluation spans a range of leading LLMs, including foundational models, those enhanced for reasoning capabilities (Xu et al., 2025), and models integrated into an iterative induction-verification pipeline designed to refine rule generation, reveal persistent challenges such as hallucination and overgeneralization, yet confirm measurable progress in rule induction. Notably, smaller-scale LLMs (3B-8B parameters) fine-tuned on our LRI Dataset demonstrate significant improvements, achieving over 76% gains in both Macro and Micro F1-scores and outperforming larger, closed-source models.

2 Related Work

2.1 Legal Reasoning in Computational Law

In the domain of computational law, research on legal reasoning has evolved along several principal paradigms. First, tasks like Legal Document Summarization (LDS) (Zhong and Litman, 2022; Shen et al., 2022; Polsley et al., 2016) and Legal Argument Mining (LAM) (Santin et al., 2023; Poudyal et al., 2020; Palau and Moens, 2009) aim to demystify legal texts by extracting structured arguments or generating layperson-friendly summaries. Another prominent direction includes Legal Question Answering (LQA) (Zhang et al., 2023b; Sovrano et al., 2020; Louis et al., 2023b) and Legal Judgment Prediction (LJP) (Zhong et al., 2020; Zhang et al., 2023a; Chalkidis et al., 2019), where systems

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

leverage existing precedents to resolve new cases, operating within deductive frameworks that apply predefined rules to specific scenarios. Advances in NLP, particularly LLMs (Minaee et al., 2025), extend these capabilities to practical applications such as automated legal consultation (Cui et al., 2024), contract review (Graham et al., 2023), and drafting (Wang et al., 2025). However, a critical gap persists: current research prioritizes rule application over rule discovery while human legal reasoning inherently combines deductive and inductive logic. To address this, we introduce LRI, which aims to extend computational jurisprudence beyond precedent-based reasoning towards the inductive formulation of legal rules.

2.2 Inductive Reasoning

Inductive reasoning (Heit, 2000) is a fundamental cognitive process that involves drawing general conclusions from specific observations. Cognitive science frames induction as probabilistic belief revision under the Bayesian framework (Tenenbaum et al., 2011), where learning arises from combining prior knowledge with observed data to derive posterior probabilities (Lake et al., 2015). NLP research in inductive reasoning recently shifts from task-specific architectures (Odena et al., 2021; Tian et al., 2020; Sablé-Meyer et al., 2022) to large pretrained models capable of broad inductive inference in natural language (Yang et al., 2024; Mirchandani et al., 2023; Gendron et al., 2024). LLMs equipped with extremely long context windows *(> 100k tokens)* and *thinking* ability (Wei et al., 2022; DeepSeek-AI et al., 2025a) can ingest multiple full-length cases and surface latent regularities without manual feature engineering. Consequently, legal rule discovery is evolving from static symbol manipulation to dynamic pattern extraction in free text. Crucially, this evolution provides systematic evidence against critiques positing legal reasoning as fundamentally analogy-based (Sherwin, 1999) or similarity-based (Schauer, 1987).

3 Preliminaries

3.1 Task Definition

We define Legal Rule Induction (LRI) as the task of algorithmically deriving a concise set of normative rules from a given collection of precedent cases. Formally, given a **precedent case set** $\mathcal{P} = \{p_i\}_{i=1}^{M}$ where $M \in \mathbb{N}^+$ ranges between 5 and 10 inclusive. Each case p_i within this set is a structured entity comprising four key components: a *fact description*, the *litigation process*, a *legal analysis*, and the final *judgment result*. The objective is to induce a rule set $\mathcal{R} = \{r_j\}_{j=1}^{N}$. Each rule r_j in this set comprises three elements: a *hypothetical condition*, a *behavior pattern*, and a *legal consequence*. To ensure that each induced rule reflects a generalizable normative pattern, rather than an isolated exception, we adopt a principle from frequent pattern mining (Agrawal and Srikant, 1994). Specifically, each rule r_j must apply to a strict majority of the precedent cases in $\mathcal{P}$. We define the support set of a rule r_j, denoted $\text{Supp}(r_j, \mathcal{P})$, as:

$$\text{Supp}(r_j, \mathcal{P}) = \{p_i \in \mathcal{P} \mid r_j \text{ applies to } p_i\}. \quad (1)$$

Accordingly, the coverage condition for each rule is:

$$|\text{Supp}(r, \mathcal{P})| > \frac{M}{2}. \quad (2)$$

3.2 Legal Rule

To ensure that the LRI task aligns with well-established legal reasoning and facilitates cross-cultural generalizability, we consider legal rules along three conceptual dimensions. From the perspective of doctrinal scope, this study focuses on **criminal law**, **civil law**, and their associated **procedural laws** (Dong and Zhang, 2023). Legal domains that are highly specialized or jurisdictionally narrow, such as administrative regulations or municipal by-laws, are excluded from consideration. From the perspective of normative function, each induced rule is required to instantiate one of the foundational deontic modalities recognized in legal theory: **permission** (an action is allowed), **prohibition** (an action is forbidden), or **obligation** (an action is required). More complex or context-dependent normative forms, such as conditional obligations or defeasible rules, fall outside the scope of the present formulation and are left for future work. From the perspective of normative source, legal rules are divided into two broad categories. (1) **Explicit rules** are those that are codified in statutes, regulations, or other formal legal texts. (2) **Implicit rules**, by contrast, are not directly articulated in legal provisions but are inferred through analysis of legal reasoning across precedent cases. For example, in Chinese criminal adjudication, male defendants often receive harsher sentences than female defendants for comparable offenses, even though no statute explicitly mandates such a distinction (Lao, 2022).

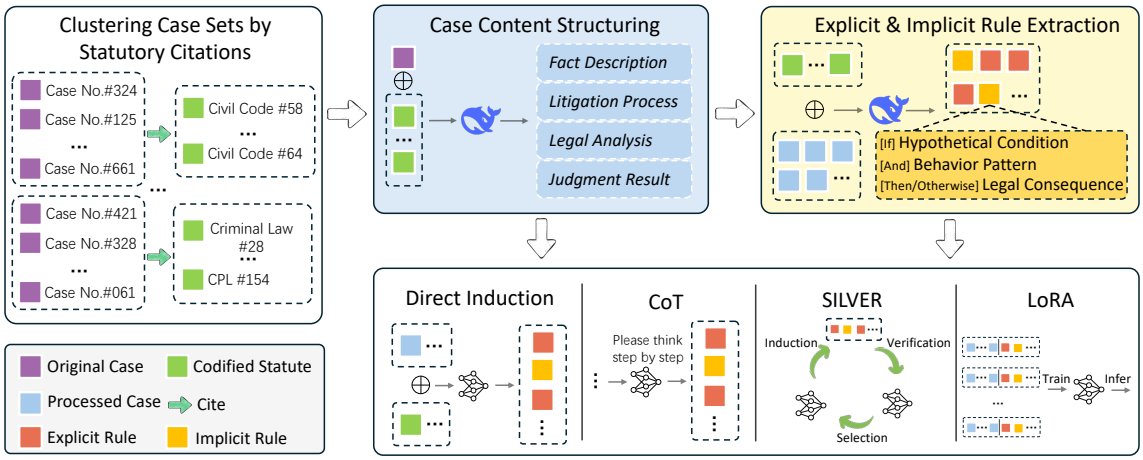

Figure 2: The overview of the **LRI-AUTO** dataset curation pipeline (for civil and criminal cases) and main methods for rule induction, including LoRA, which utilizes LRI-AUTO for tuning and the **LRI-GOLD** dataset for testing.

3.3 Inductive Reasoning Pipeline

In the main experiments, we consider four training-free pipelines in inductive reasoning:

Direct Induction This pipeline employs LLMs to generate normative rules directly from the provided case texts using a single-step prompting strategy. Following (Zheng et al., 2025), we consider the direct output of LLMs in this manner as a form of baseline inductive inference.

Chain-of-Thought (CoT) CoT prompting (Wei et al., 2022) operationalizes a more deliberative, multi-step reasoning process. It guides the LLM to decompose the rule induction task into intermediate analytical stages (*e.g.*, identifying common factual patterns, discerning judicial reasoning, and then formulating a rule).

Long Chain-of-Thought Long-CoT refers to the phenomenon where Large Reasoning Models (LRMs), such as o1 (Jaech et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025a), spontaneously generate extended chains of reasoning before answering complex questions.

SILVER To further advance rule induction for LRI, we propose **SI**mp**L**y Iterative Induction and **VER**ification **(SILVER)**, which implements an induction–verify–update loop (Qiu et al., 2024). As detailed in Algorithm 1, the process commences with an initial set of rules induced from the case sets. Subsequently, SILVER alternates between another two core stages as detailed in Appendix C.4: (i) verifying each candidate rule against the case set to determine if it surpasses the predefined majority-support threshold, and (ii) re-inducing fresh candidate rules to address aspects of the cases not adequately covered by the already verified ones. This cycle repeats until no new high-support rules are found or a maximum iteration count is reached.

4 Legal Rule Induction Dataset

In this section, we present the Legal Rule Induction dataset curation pipeline, as detailed in Figure 2, and provide dataset statistics.

4.1 Corpus and Clustering

Chinese legal cases typically specify cited legal article numbers, enabling large-scale automated clustering of case sets sharing common legal bases. We collect over 9 million criminal/civil cases from China Judgments Online (CJO) and their contemporaneous legal provisions to ensure citation consistency (see Appendix B.1). Using regex, we extract all legally cited provisions of these cases from four core Chinese legal codes: the ***Criminal Law***, the ***Civil Code***, the ***Criminal Procedure Law***, and the ***Civil Procedure Law***. Then, cases citing identical legal provisions are automatically clustered into the same case sets, and the set size distribution is depicted in Figure 7.

4.2 Case Content Structuring

Original documents contain regional formatting inconsistencies and sensitive information such as court names, personal identifiable information (PII), and legal article texts. To isolate the core case content and legal citation for each case $p \in \mathcal{P}$, we employ the DeepSeek-R1 model (DeepSeek-AI et al., 2025a) for content structuring with anonymisation. Building on (Huang et al., 2024), we iden-

	# Train	# Test	# Gold
Case Sets	4,552	569	216
Civil Case Sets	2,847	347	108
Criminal Case Sets	1,705	222	108
Cases	33,797	4,291	1,620
Civil Cases	21,068	2,601	810
Criminal Cases	12,729	1,690	810
Rules	26,372	3,278	1,132
Explicit Rules	15,608	1,933	711
Implicit Rules	10,764	1,345	421
Avg Case Length	569.5	567.1	569.0
Avg Rule Per Case Set	5.79	5.76	5.24
Annotation	-	-	✔

Table 1: Statistics for automatically constructed LRI-AUTO (Train/Test) and expert-annotated LRI-GOLD.

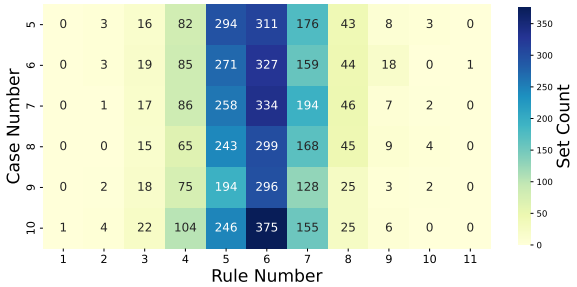

Figure 3: Distribution of rule set sizes across case numbers in the LRI Dataset.

tify and extract four key components from the court documents for each case. We replace their *relevant law* section from the structured case content with the litigation process (also known as procedural history) to avoid exposing legal articles/charges directly while ensuring LLMs access complete procedural context during rule induction. Sensitive data (*e.g.*, names → "Defendant A", locations → "City C") is anonymised with generic substitutes, preserving demographic details (age/gender/occupation) where pertinent. Full implementation protocols are in Appendix B.2.

4.3 Explicit and Implicit Rule Extraction

Legal provisions $\mathcal{S}_p$, defined as the set of statutory articles cited by the case set $\mathcal{P}$, associated with each case set $\mathcal{P}$ (Section 4.1), are unsuitable as direct ground truth rules for LRI. Firstly, $\mathcal{S}_p$ often contains specific charges or offence names, skewing LRI towards statutory retrieval instead of rule induction. Secondly, cases may not use all parts of cited provisions, as articles often have multiple sub-clauses (*e.g.*, a case set might only pertain to one paragraph of a multi-paragraph article like Article 1079, PRC Civil Code, despite the entire article being cited). Therefore, for each case set $\mathcal{P}$, DeepSeek-R1 is used to derive the two rule categories: (1) **Explicit rules** r_{exp}: Rules directly from $\mathcal{S}_p$ applicable to all cases in $\mathcal{P}$, excluding specific charges/offense names. (2) **Implicit rules** r_{imp}: Rules reflecting judicial practices or societal norms, not explicit in $\mathcal{S}_p$. Rule extraction prompts and methodologies are detailed in Appendix B.3.

4.4 Case Set Postprocessing

Rule Element Integrity Filter To ensure rule completeness, case sets are filtered if their corresponding rules, as extracted by DeepSeek-R1, lack essential elements in the ***hypothetical condition***, ***behavior pattern*** (including action type), or ***legal consequence***. This addresses potential omissions due to DeepSeek-R1 limitations, like hallucination or inconsistent instruction following.

Rule Applicability Filter A filtering step is applied to refine the rule sets: explicit rules r_{exp} are retained only if they demonstrate 100% applicability across all cases within their respective set $\mathcal{P}$. Implicit rules r_{imp} are retained only if their applicability, as initially assessed, exceeds the 50% threshold within their set.

Set Size Filter To manage the solution space for rule induction and constrain model input context, sets are filtered to retain those with over 5 cases. Sets exceeding 10 cases are randomly sampled down to 10. This results in final case sets $\mathcal{P} = \{p_1, p_2, \ldots, p_M\}$ containing 5 to 10 cases.

4.5 LRI Dataset Collection and Annotation

Following DeepSeek-R1 response collection and several filters, the LRI-AUTO dataset is constructed for model training. This involves uniformly sampling approximately 1,000 instances from case set collections, categorized by the number of cases per set (ranging from 5 to 10). Each sampled instance comprises a case set $\mathcal{P}$ and its corresponding rule set $\mathcal{R}$. For robust evaluation, we further construct the LRI-GOLD test set by uniformly sampling a smaller, balanced subset of criminal and civil cases. Three annotators independently extract and induce rule sets for this subset, following strict annotation guidelines. Details of the guidelines, annotator backgrounds, and inter-annotator agreement are provided in Appendix B.4.

4.6 Dataset Statistics and Expert Analysis

Table 1 provides detailed LRI dataset statistics. The LRI-AUTO dataset comprises 5,121 case sets (to-

Method	Model	Rule Type		Rule Level		Set Level	
		Exp-Rec	Imp-Rec	Mic-Pre	Mic-F1	Mac-Pre	Mac-F1
LLMs (Direct)	GPT-4o-mini	45.99	29.22	57.25	46.92	58.41	46.86
	GPT-4o	55.56	27.79	**71.81**	55.50	**72.65**	54.53
	Gemini-2.5-Flash	73.00	37.77	61.14	60.51	60.74	58.96
	Llama-4-Scout	47.26	25.18	58.47	46.82	60.87	45.85
	Llama-4-Maverick	48.10	23.04	60.39	47.23	59.68	45.59
	Qwen-2.5-72b	62.17	42.76	58.24	56.55	60.10	55.50
	Qwen-Max	60.76	43.47	61.32	57.61	60.05	56.14
	DeepSeek-V3-0324	66.10	**47.74**	62.83	61.00	61.66	59.27
	Claude-3.5-Sonnet	59.63	35.39	70.74	59.01	70.73	58.40
	Claude-3.7-Sonnet	**74.68**	42.99	70.92	**66.67**	70.23	**65.22**
LLMs (CoT)	GPT-4o-mini	41.49	15.68	67.98	43.42	67.69	42.53
	GPT-4o	41.63	14.49	**80.95**	45.39	**78.36**	43.72
	Gemini-2.5-Flash	68.21	28.50	73.78	61.99	74.14	60.51
	Llama-4-Scout	45.85	17.10	71.97	47.24	75.25	46.41
	Llama-4-Maverick	41.49	14.73	72.71	43.99	72.03	41.90
	Qwen-2.5-72b	44.02	21.14	68.37	46.74	70.26	44.32
	Qwen-Max	54.29	24.47	72.44	54.12	72.76	52.95
	DeepSeek-V3-0324	61.88	32.07	69.70	58.76	71.87	56.65
	Claude-3.5-Sonnet	54.15	26.60	74.18	55.16	73.80	54.69
	Claude-3.7-Sonnet	**70.89**	**37.77**	75.77	**66.07**	76.66	**65.17**
LRMs (Long-CoT)	o3-mini	46.41	13.78	**83.08**	48.53	**84.04**	48.76
	Gemini-2.5-Flash	72.01	31.83	70.22	62.96	70.58	61.37
	Deepseek-R1	62.87	41.57	74.31	63.18	74.45	61.38
	Claude-3.7-Sonnet	**75.39**	**43.94**	68.02	**65.78**	67.94	**64.33**
	Grok-3-mini	42.76	13.30	70.73	43.88	71.18	43.94
LLMs (SILVER)	GPT-4o-mini	68.92	38.48	58.01	57.80	55.13	54.95
	Gemini-2.5-Flash	**88.19**	43.71	62.15	**66.56**	60.42	**63.82**
	Llama-4-Scout	64.14	29.93	**63.89**	55.70	**63.19**	54.87
	Qwen-2.5-72b	81.01	52.02	57.11	63.00	52.68	57.82
	DeepSeek-V3-0324	84.81	**64.77**	56.99	64.73	51.88	59.90

Table 2: Performance (%) on the LRI-GOLD benchmark across four baselines. **Exp-Rec** and **Imp-Rec** denote Micro Recall on explicit and implicit rules. We **bold** the best and underline the second-best results in each baseline.

talling 38,088 cases and 29,650 rules), with 4,552 sets for training and 569 for testing. The criminal-to-civil case ratio in LRI-AUTO (approx. 1:1.6) reflects the original CJO corpus distribution. The LRI-GOLD test set contains 108 criminal and 108 civil case sets. Figure 3 illustrates the numerical distribution of cases and rules per set across the dataset. A manual audit conducted on 100 randomly selected LRI-AUTO sets, utilizing criteria specified in Table 7, confirmed that the vast majority of rules correctly apply to their respective case sets.

5 Experiments

In this section, we assess LLMs performance on the LRI-GOLD benchmark and demonstrate how LRI-AUTO enhances legal rule induction in smaller models through parameter-efficient adaptation.

5.1 Experimental Settings

Baseline Methods As discussed in Section 3.3, we compare several approaches: Direct Induction (zero-shot prompting), CoT (prompting with "think step by step"), Long-CoT (reasoning before responding), SILVER (an automatic induction-verification pipeline), and fine-tuning on LRI-AUTO for small LLMs (3B-8B). Detailed prompt templates for the above methods are provided in Appendix C.

Models We conduct experiments on three types of LLMs as depicted in Appendix C.1: (1) LLMs: direct inference without thinking before response, (2) LRMs as detailed in Section 3.3, equipped with Long-CoT ability and think before response, (3) Small-size LLMs, whose parameter number is below or equal to 8 billion.

Evaluation Metrics We assess induced rule quality and correctness using DeepSeek-V3 (DeepSeek-AI et al., 2025b) as an automated judge, employing two complementary perspectives: (1) **Rule Level (Micro) Evaluation**: This metric assesses all induced rules individually, disregarding their case set origins, to emphasize overall rule correctness

Method	Model	Rule Type		Rule Level		Set Level	
		Exp-Rec	Imp-Rec	Mic-Pre	Mic-F1	Mac-Pre	Mac-F1
LLMs	Llama-3.2-3B	24.91	9.97	19.13	19.21	17.99	17.89
	Ministral-3B	41.49	21.14	36.54	35.18	37.01	32.83
	Qwen-2.5-7B	**58.23**	**33.73**	**56.45**	**52.53**	**57.97**	**50.75**
	Ministral-8B	48.66	21.38	44.81	41.43	45.81	39.78
LLMs + LoRA	Llama-3.2-3B	83.54	51.07	70.47	70.96	67.63	68.31
	Ministral-3B	78.96	38.05	60.79	62.25	55.61	58.48
	Qwen-2.5-7B	**83.68**	56.06	70.07	71.70	66.73	68.65
	Ministral-8B	83.31	**58.00**	**72.47**	**73.18**	**70.19**	**70.73**

Table 3: Performance (%) of four small-sized LLMs and their performance after LoRA fine-tuning on LRI-AUTO.

(akin to micro-averaging). It is calculated as:

$$\text{Mic-F1} = \frac{2 \cdot \text{Mic-Pre} \cdot \text{Mic-Rec}}{\text{Mic-Pre} + \text{Mic-Rec}}, \quad (3)$$

where Mic-Pre (micro-precision) is the total number of correctly predicted rules divided by the total number of predicted rules across all case sets, and Mic-Rec (micro-recall) is the total number of correctly predicted rules divided by the total number of gold-standard rules across all case sets. (2) **Set Level (Macro) Evaluation**: This metric evaluates performance on a per-case-set basis, treating each as an independent unit and averaging their F1 scores:

$$\text{Mac-F1} = \frac{1}{N_{\text{sets}}} \sum_{i=1}^{N_{\text{sets}}} \text{F1}(\mathcal{R}_i^{\text{pred}}, \mathcal{R}_i^{\text{gold}}), \quad (4)$$

where $\mathcal{R}_i^{\text{pred}}$ and $\mathcal{R}i^{\text{gold}}$ are the predicted and gold-standard rule sets for the i-th case set, and N_{sets} is the total number of precedent case sets. To assess the **reliability** of DeepSeek-V3 as an automated judge, we provide a detailed comparison with human evaluation results in Appendix C.3.

5.2 Main Evaluation

Performance Comparison across Inductive Pipelines Analysis of Table 2 and Figure 4 reveals distinct performance characteristics of different inductive pipelines. CoT prompting generally enhances precision at the cost of recall, leading to a slight decrease in F1 scores for most LLMs compared to Direct Induction. For instance, GPT-4o's (Hurst et al., 2024) Micro-Precision rises from 71.81% to 80.95%, while its explicit rule recall drops from 55.56% to 41.63%. Exceptions like Gemini-2.5-Flash (Mic-F1 +1.48%) suggest model-specific benefits. Long-CoT presents varied outcomes: Gemini-2.5-Flash (Deepmind, 2025) (Long-CoT) improves precision (Mic-Pre +9.08%) and Mic-F1 (+2.45%) over its direct counterpart,

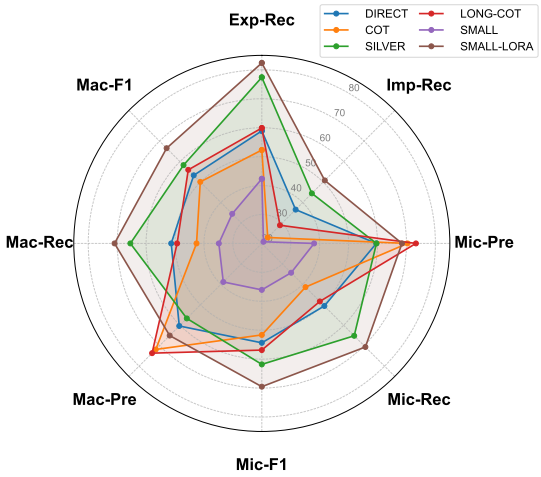

Figure 4: Scores (%) of different baselines. For the Direct, CoT, and SILVER baselines, only the five LLMs common to all three are considered.

albeit with reduced recall. Conversely, Claude-3.7-Sonnet (Anthropic, 2025) (Long-CoT) showed increased recall (Exp-Rec +0.71%) but lower precision (Mic-Pre -2.90%) and Mic-F1 (-0.89%). This indicates that extended reasoning contexts affect the precision-recall balance differently across models. The SILVER pipeline consistently yields superior performance, primarily through substantial recall improvements across models (*e.g.*, Gemini-2.5-Flash Exp-Rec increased from 73.00% to 88.19%), leading to higher F1 scores (*e.g.*, DeepSeek-V3-0324 Mic-F1 improved from 61.00% to 64.73%). This underscores the efficacy of SILVER's multi-turn induction and verification.

Efficacy of LRI-AUTO Table 3 demonstrates the effectiveness of LRI-AUTO dataset in enhancing small LLMs (3B-8B) performance. Initially, these models show limited capabilities (*e.g.*, Llama-3.2-3B Mic-F1 19.21%). However, LoRA fine-tuning (Han et al., 2024) on LRI-AUTO yields substantial gains across all metrics for all four tested small LLMs. For example, Mic-F1 of

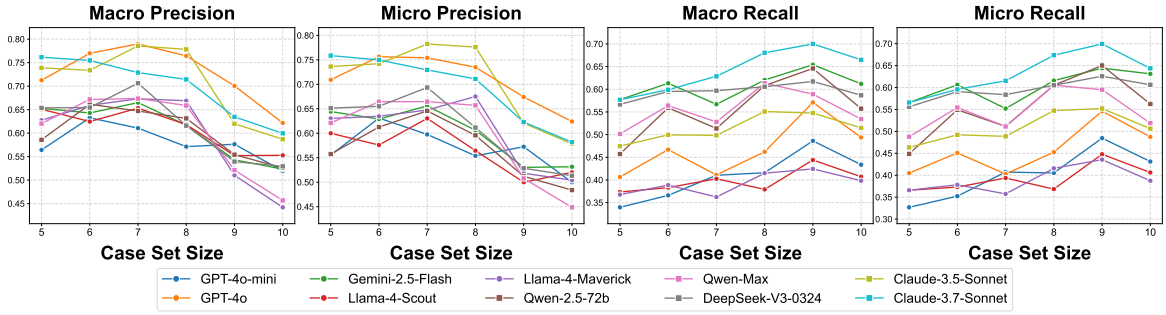

Figure 5: Performance trends of Direct Induction of ten LLMs across varying case set sizes.

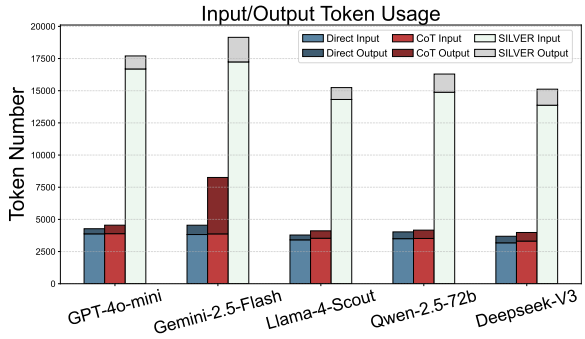

Figure 6: Comparison of token usage (Input & Output) for different LLMs under three different baselines.

Llama-3.2-3B surged to 70.96%. Notably, the fine-tuned Ministral-8B (+LoRA) achieves a Mic-F1 of 73.18% and Mac-F1 of 70.73%. This performance surpasses several larger proprietary models under Direct Induction prompting (Table 2), such as Gemini-2.5-Flash (Direct Mic-F1 60.51%) and Claude-3.7-Sonnet (Direct Mic-F1 66.67%).

5.3 Further Discussion

Explicit and Implicit Rule In the LRI evaluation phase, LLMs are not informed whether rules are explicit (directly from statutes) or implicit. We observe consistently higher recall for explicit rules. We attribute this disparity to two primary factors. First, explicit rules are designed to be present across all cases within a given case set, which inherently increases their discoverability and ease of extraction by the models. Second, even when specific crime names are masked, LLMs with pre-existing knowledge of Chinese law (from their training data) (Fei et al., 2023) tend to exhibit greater sensitivity to the linguistic patterns characteristic of these explicit, statute-like rules. Conversely, implicit rules, requiring deeper inference, are harder to identify. This suggests that performance on implicit rules may better reflect an LLM's ability to generalize in unfamiliar legal domains.

Set Size Sensitivity As shown in Figure 5, LLM performance in legal rule induction varies with case set size. Generally, increasing the number of input cases leads to lower precision but higher recall. This is likely due to overgeneration of broad or less accurate rules, improving coverage (recall) but reducing accuracy (precision). With fewer cases, it's harder for models to detect shared patterns, leading to lower recall. Claude-3.7-Sonnet and DeepSeek-V3-0324 show stable performance across different sizes, while GPT-4o-mini and Llama-4-Maverick degrade more sharply, indicating difficulties in balancing abstraction and specificity.

Token Usage Figure 6 reveals that the SILVER pipeline incurs the highest token consumption due to its iterative multi-turn architecture. Gemini-2.5-Flash and DeepSeek-V3 are particularly token-intensive under SILVER. Direct Induction prompting is the most token-efficient but, as noted in Section 5.2, typically results in lower performance. The CoT strategy moderately increases token output compared to Direct Induction, but this often does not translate into commensurate F1 score improvements, potentially diminishing its cost-effectiveness. These observations underscore the critical trade-off between computational efficiency and reasoning depth in practical applications.

6 Conclusion

This paper formalizes Legal Rule Induction (LRI) as the task of distilling rules from analogous cases and introduces the first benchmark comprising LRI-AUTO for tuning and expert-annotated LRI-GOLD for evaluation. Our experiments demonstrate that while leading LLMs initially struggle with this complex extraction task, training on our dataset significantly improves their rule induction capabilities. This work establishes a foundation for LRI in the LLM era and addresses a critical gap in computational legal reasoning research.

Limitations

Our research, conducted within the Chinese legal system, exclusively utilizes Chinese-language legal cases and rules. This grounding in a specific jurisdiction and language introduces limitations: the models may exhibit a bias towards the Chinese legal framework, potentially restricting their direct generalizability to other legal systems without adaptation, and their performance in multilingual contexts remains unassessed. Furthermore, this work did not investigate the utility of our legal rule induction methods on downstream applications such as legal information retrieval (Sansone and Sperlí, 2022), judgment prediction (Cui et al., 2022), or question answering (Martinez-Gil, 2023); exploring this efficacy presents a significant avenue for future research.

Ethics Statement

The source materials for our dataset are derived from a publicly accessible platform, China Judgments Online, which is the largest legal judgment database in China and is widely used in academic research (Huang et al., 2024; Xiao et al., 2018). Any specific legal provisions and personally identifiable information (PII) encountered are rigorously anonymised during the dataset construction process. Human annotators involved in the project are compensated at a rate of 15 USD per hour, a figure that exceeds the prevailing minimum wage in China. To the best of our knowledge, this work adheres to all relevant open-source agreements and does not pose risks of information leakage or other ethical hazards. Beyond data privacy and legality, we acknowledge that inducing legal rules from judicial texts may entail broader societal and ethical implications. Models trained on judicial texts may reflect or amplify historical biases (e.g., related to gender or region) and could misrepresent complex legal reasoning if used without expert oversight. The LRI datasets and models are therefore released solely for academic research and methodological evaluation, with an emphasis on responsible use and further study of fairness, transparency, and accountability in computational law.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, page 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mistral AI. Ministral 8b instruct.

Anthropic. 2024. Claude 3.5 sonnet system card.

Anthropic. 2025. Claude 3.7 sonnet system card.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? *Preprint*, arXiv:2302.06100.

Scott Brewer. 2013. *Precedents, Statutes, and Analysis of Legal Concepts: Interpretation*. Routledge.

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.

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

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *Preprint*, arXiv:2204.04859.

Deepmind. 2025. Gemini 2.5 flash.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, 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, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan,

Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

John Dickinson. 1931. Legal rules: their function in the process of decision. *University of Pennsylvania Law Review and American Law Register*, 79(7):833–868.

Xiaobo Dong and Yafang Zhang. 2023. Procedural laws. In *On Contemporary Chinese Legal System*, pages 311–339. Springer.

William O Douglas. 1949. Stare decisis. *Columbia Law Review*, 49(6):735–758.

Melvin A. Eisenberg. 2022. *Legal Reasoning*. Cambridge University Press.

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

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2024. Large language models are not strong abstract reasoners. *Preprint*, arXiv:2305.19555.

S Georgette Graham, Hamidreza Soltani, and Olufemi Isiaq. 2023. Natural language processing for legal document review: categorising deontic modalities in contracts. *Artificial Intelligence and Law*, pages 1–22.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built

benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Preprint*, arXiv:2403.14608.

Evan Heit. 2000. Properties of inductive reasoning. *Psychonomic bulletin & review*, 7:569–592.

Oliver Wendell Holmes Jr. 2020. *The common law*. Routledge.

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

Wanhong Huang, Yi Feng, Chuanyi Li, Honghan Wu, Jidong Ge, and Vincent Ng. 2024. CMDL: A large-scale Chinese multi-defendant legal judgment prediction dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5895–5906, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Grant Lamond. 2005. Do precedents create rules? *Legal Theory*, 11(1):1–26.

Jiaqi Lao. 2022. Extra-legal factors in sentencing and the reform of sentencing standardization. *Chinese Journal of Criminal Law*, (2):91–107.

Lei Lei. 2013. The logical structure of legal rules. *Legal Studies*, 35(1):66–86.

Edward H Levi. 2013. *An introduction to legal reasoning*. University of Chicago Press.

Yuqi Liu and Yan Zheng. 2025. Improving similar case retrieval ranking performance by revisiting ranksvm. *Preprint*, arXiv:2502.11131.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023a. Finding the law: Enhancing statutory article retrieval via graph neural networks. *Preprint*, arXiv:2301.12847.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023b. Interpretable long-form legal question answering with retrieval-augmented large language models. *Preprint*, arXiv:2309.17050.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

John Merryman and Rogelio Pérez-Perdomo. 2018. *The civil law tradition: an introduction to the legal systems of Europe and Latin America*. Stanford University Press.

Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Ministral. Ministral 3b instruct.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large language models as general pattern machines. *Preprint*, arXiv:2307.04721.

Augustus Odena, Kensen Shi, David Bieber, Rishabh Singh, Charles Sutton, and Hanjun Dai. 2021. Bustle: Bottom-up program synthesis through learning-guided exploration. *Preprint*, arXiv:2007.14381.

OpenAI. Openai o3-mini.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. Casesummarizer: A system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 258–262, Osaka, Japan. The COLING 2016 Organizing Committee.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and

Xiang Ren. 2024. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *Preprint*, arXiv:2310.08559.

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. *Preprint*, arXiv:2412.15115.

Mathias Sablé-Meyer, Kevin Ellis, Josh Tenenbaum, and Stanislas Dehaene. 2022. A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, 139:101527.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

Piera Santin, Giulia Grundler, Andrea Galassi, Federico Galli, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2023. Argumentation structure prediction in cjeu decisions on fiscal state aid. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 247–256, New York, NY, USA. Association for Computing Machinery.

Frederick Schauer. 1987. Precedent. *Stanford Law Review*, pages 571–605.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Preprint*, arXiv:2206.10883.

Emily Sherwin. 1999. A defense of analogical reasoning in law. *U. Chi. L. Rev.*, 66:1179.

Francesco Sovrano, Monica Palmirani, and Fabio Vitali. 2020. Legal knowledge extraction for knowledge graph based question-answering. In *Legal knowledge and information systems*, pages 143–153. IOS Press.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Lucas Y. Tian, Kevin Ellis, Marta Kryven, and Joshua B. Tenenbaum. 2020. Learning abstract structure for drawing by efficient motor program induction. *Preprint*, arXiv:2008.03519.

Steven H. Wang, Maksim Zubkov, Kexin Fan, Sarah Harrell, Yuyang Sun, Wei Chen, Andreas Plesner, and Roger Wattenhofer. 2025. Acord: An expert-annotated retrieval dataset for legal contract drafting. *Preprint*, arXiv:2501.06582.

Thomas Glyn Watkin. 2017. *An historical introduction to modern civil law*. Routledge.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Zhang Wenxian, Li Long, Zhou Wangsheng, Zheng Chengliang, and Xu Xianming. 2018. *Jurisprudence (5th Edition)*. Higher Education Press, Beijing.

xAI. Grok 3 beta — the age of reasoning agents.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *Preprint*, arXiv:2501.09686.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2024. Language models as inductive reasoners. *Preprint*, arXiv:2212.10923.

Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023a. Contrastive learning for legal judgment prediction. *ACM Trans. Inf. Syst.*, 41(4).

Weiqi Zhang, Hechuan Shen, Tianyi Lei, Qian Wang, Dezhong Peng, and Xu Wang. 2023b. GLQA: A generation-based method for legal question answering. In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pages 1–8. IEEE.

Tianshi Zheng, Jiayang Cheng, Chunyang Li, Haochen Shi, Zihao Wang, Jiaxin Bai, Yangqiu Song, Ginny Y. Wong, and Simon See. 2025. Logidynamics: Unraveling the dynamics of logical inference in large language model reasoning. *Preprint*, arXiv:2502.11176.

Haoxiang Zhong, Yuzhong Wang, Cunchao Tu, T. Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *AAAI Conference on Artificial Intelligence*.

Yang Zhong and Diane Litman. 2022. Computing and exploiting document structure to improve unsupervised extractive summarization of legal case decisions. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 322–337, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Legal Rule and Jurisprudence Foundation

The foundational structure of a legal rule is commonly understood as an **if-then** conditional statement. This can be formally expressed in logical notation as:

$$\textit{Condition} \rightarrow \textit{Consequence} \tag{5}$$

In civil law jurisdictions, legal rules are typically explicitly stipulated and codified within statutes. For instance, specific articles within the *Civil Code of the People's Republic of China*[2] or the French *Code Civil (Napoleonic Code)*[3] clearly delineate such rules, providing a primary source for legal reasoning. Conversely, common law rules are specific legal norms established by courts through precedent. Common law reasoning is also rule-based (Eisenberg, 2022), applying these court-derived rules to case facts. The rule a precedent establishes is its *holding*—the explicit legal principle stated by the court as governing the case, which forms binding law. Other judicial statements within a precedent, known as *dicta*, are not binding but may possess persuasive influence. This paper, focusing on Chinese legal reasoning, adopts the "three-element theory" from Chinese jurisprudence (Wenxian et al., 2018). This theory structures a rule with a: **hypothetical condition**, **behavior pattern**, and **legal consequence** logically represented as:

$$\text{Hypothetical Condition} \wedge \text{Behavior Pattern} \rightarrow \text{Legal Consequence} \tag{6}$$

An alternative, the new "two-element theory", posits rules as **constituent elements** and **legal consequences**. It suggests the behavior pattern is integrated within these two, aiming for a unified structure for various rule types. However, we find that LLMs struggle to accurately interpret the new two-element theory, often producing erroneous outputs. Therefore, for reliability in this study, we utilize the more widely understood and LLM-compatible three-element theory.

[2] https://english.www.gov.cn/archive/lawsregulations/202012/31/content_WS5fedad98c6d0f72576943005.html

[3] https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006070721/

B Details of LRI Dataset

B.1 China Judgments Online (CJO)

This study utilizes CJO case data and legal article versions from 2021 due to two key factors. Primarily, the substantial public availability of 2021 case datasets makes them highly suitable for clustering purposes. Furthermore, the enactment of the Chinese Civil Code in 2020, which integrates seven distinct legal domains (General Provisions, Property Rights, Contracts, Personality Rights, Marriage and Family, Inheritance, and Tort Liability) previously governed by separate statutes, streamlines the process of systematic legal article extraction by avoiding the increased labor costs associated with mapping article numbers and content from a fragmented pre-2021 legal landscape.

B.2 Prompt of Case Content Structuring

To facilitate a comprehensive presentation of a legal case's factual background, procedural history, judicial analysis, and adjudicated outcome, while concurrently ensuring the anonymisation of sensitive entities and the abstraction of specific legal article numbers and their textual content, we formulate the prompt delineated in Table 8. This prompt utilizes the original legal case document and the content of cited legal provisions as input, which are subsequently processed by the DeepSeek-R1 model (DeepSeek-AI et al., 2025a).

B.3 Prompt of Rule Extraction

Using the structured case contents and the content of cited legal provisions, we employ the prompts detailed in Table 9 and Table 10. Explicit and implicit rules are subsequently collected from the responses generated by DeepSeek-R1.

B.4 Human Annotations

To ensure high-quality rule extraction, we develop a concise and clear annotation guideline, as shown in Table 5, based on the prompt designs in Table 9 and Table 10. The annotation task is carried out by three trained annotators, all graduate students in law from China, possessing a solid theoretical background in both Chinese criminal and civil law. Before commencing the full annotation process, we conduct a pilot annotation experiment to assess inter-annotator similarity. Specifically, the three annotators independently annotate 67 explicit and 39 implicit rules across 20 case sets. We then compute pairwise Jaccard similarities to evaluate the

	Annotator 1	Annotator 2	Annotator 3
Explicit Rules			
Annotator 1	—	0.870	0.887
Annotator 2	0.870	—	0.901
Annotator 3	0.887	0.901	—
Implicit Rules			
Annotator 1	—	0.553	0.571
Annotator 2	0.553	—	0.574
Annotator 3	0.571	0.574	—

Table 4: Pairwise Jaccard similarities among annotators for explicit and implicit rules

consistency of extracted rule sets among annotators. The results are presented in Table 4. For explicit rules, the pairwise Jaccard similarities are relatively high (0.870–0.901), indicating strong agreement among annotators. In contrast, for implicit rules, the similarities are lower (0.553–0.574), suggesting that identifying implicit legal rules involves greater subjectivity and interpretive variation.

B.5 Statistics

Figure 7 illustrates the original case set size distribution. Figure 8 depicts the case length distribution within the LRI dataset, with most cases ranging from 400 to 600 Chinese characters in length. Furthermore, Figure 9 presents the distribution of rules per case set in the LRI dataset. We observe that the number of rules per case set in LRI-GOLD is slightly lower than in LRI-AUTO.

C Implementation Details

C.1 Model Details

In our experiments, we evaluate a total of 19 LLMs, categorized as follows:

LLMs GPT-4o-mini(Hurst et al., 2024), GPT-4o (Hurst et al., 2024), Gemini-2.5-Flash (Deepmind, 2025), Llama-4-Scout (Meta, 2025), Llama-4-Maverick (Meta, 2025), Qwen-2.5-72b (Qwen et al., 2025), Qwen-Max (Qwen et al., 2025), DeepSeek-V3-0324 (DeepSeek-AI et al., 2025b), Claude-3.5-Sonnet (Anthropic, 2024), Claude-3.7-Sonnet (Anthropic, 2025).

LRMs DeepSeek-R1 (DeepSeek-AI et al., 2025a), o3-mini (OpenAI), Grok-3-mini (xAI), Claude-3.7-Sonnet:Thinking (Anthropic, 2025), Gemini-2.5-Flash:Thinking (Deepmind, 2025).

1. **Legal Rule Categories:** Use only one of the following: (1) Criminal (2) Civil (3) Procedural (Litigation Procedure)
2. **Legal Rule Structure:** Each rule must include:
   (a) **Hypothetical Condition** – the context and subject.
   (b) **Behavior Pattern** – classified as:
   - Permissive: "may", "is allowed to".
   - Obligatory: "must", "shall".
   - Prohibitive: "must not", "is prohibited".

   (c) **Legal Consequence** – result of compliance or violation.
3. **Rule Types:**
   (a) **Explicit Rules:**
   - Must be derived directly from cited laws.
   - Must apply to **all** cases in the set.
   - No article numbers or direct quotes.

   (b) **Implicit Rules:**
   - Inferred from **majority** (above 50%) of cases.
   - Reflect judicial discretion or practice.
4. **Formatting Requirements:**
   - Follow the logic: `If [condition], and [behavior], then/otherwise [consequence].`
   - Avoid redundancy; merge similar rules.
   - Avoid omissions; especially for cited laws.
   - Replace legal terms with plain language.
5. **Metadata:** Count applicable cases for each rule.
6. **Output Format:** Use JSON with keys: Explicit Rule, Implicit Rule.

Table 5: Annotation guideline for legal rule induction

Small LLMs Llama-3.2-3B (Meta), Ministral-3B (Ministral)/Ministral-8B (AI), Qwen-2.5-7B (Qwen et al., 2025).

All LLMs and LRMs are accessed via the OpenRouter API[4], while the small LLMs are obtained from Hugging Face[5].

C.2 LoRA Training Setting

Four open-source language models, each supporting at least an 8K token input context, are selected for instruction-tuning on the GPU with 80GB of VRAM and 1,513 TFLOPS. Specifically, these models are fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2021) for parameter efficiency. For LoRA, both the rank and alpha are set

[4] https://openrouter.ai/
[5] https://huggingface.co/

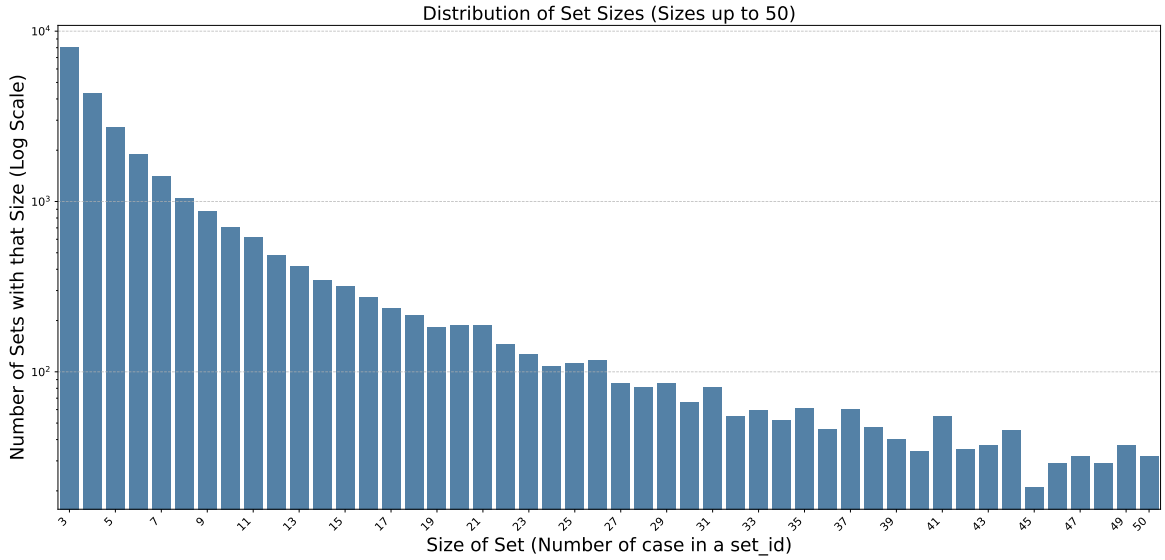

Figure 7: Original case set size distribution before re-sampling.

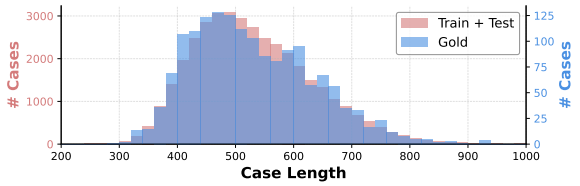

Figure 8: Case length distribution in LRI dataset.

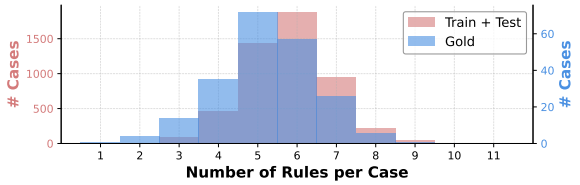

Figure 9: Rule number per case set distribution in LRI dataset.

to 8. All models are trained for 3 epochs, and their final checkpoints are used for evaluation. Other training parameters include a batch size of 8, a learning rate of 1e-4, a cutoff length of 8192 tokens, and a warmup ratio of 0.1. The training time for each model ranges from 4 to 8 hours.

C.3 LLM-as-a-Judge

Given that a single legal rule can be expressed in various linguistic forms, standard automatic evaluation metrics such as exact match, ROUGE, and BLEU are unsuitable for assessing rule induction. Consequently, we used DeepSeek-V3 as an LLM-as-a-Judge to evaluate the logical equivalence between induced rules and ground-truth rules. The prompts utilized for this evaluation are detailed in Table 16. To check the quality of the LLM-as-a-Judge, we manually reviewed the judgments made by DeepSeek-V3 on 114 rules. These rules are sampled by selecting 3 rules from each of the 38 distinct models and settings from the test phase. The outcomes of this quality assessment, presented in Table 6, show that DeepSeek-V3 performs with high accuracy on this type of classification task.

Algorithm 1 The pipeline of simply iterative induction and verification

Require: Case set $\mathcal{P}$, Threshold τ (*e.g.*, 50%), Maximum iterations max_iter
Ensure: Final rule set $\mathcal{R}_{\text{final}}$

$\mathcal{R}_{\text{final}} \leftarrow \varnothing$
$\mathcal{R}_{\text{cand}} \leftarrow \textsc{InduceInitialRules}(\mathcal{P})$
$iter \leftarrow 0$
while $iter <$ max_iter **do**
  $\mathcal{R}_{\text{verified}} \leftarrow \textsc{VerifyAndSelect}(\mathcal{R}_{\text{cand}}, \mathcal{P}, \tau)$
  if $\mathcal{R}_{\text{verified}} = \varnothing$ **then**
    break
  end if
  $\mathcal{R}_{\text{final}} \leftarrow \mathcal{R}_{\text{final}} \cup \mathcal{R}_{\text{verified}}$
  $\mathcal{R}_{\text{cand}} \leftarrow \textsc{InduceNewRules}(\mathcal{P}, \mathcal{R}_{\text{final}})$
  $iter \leftarrow iter + 1$
end while
return $\mathcal{R}_{\text{final}}$

C.4 Prompt of SILVER

The SILVER workflow includes three main stages. First, an initial round of legal rule induction is performed using the prompt specified in Table 11. Second, a legal rule verification step, utilizing the prompt in Table 15, checks if each induced rule

applies to a majority (over 50%) of cases within the given case set. Third, a subsequent round of inducing new rules from the case set is conducted, guided by the prompts detailed in Table 13 and Table 14. This stage uses the legal case set and the rule set generated in the preceding round as input.

C.5 Prompt of Direct Induction (Evaluation Phase)

For the evaluation phase of legal rule induction, we design the prompt shown in Table 11 and Table 12. The input for this prompt consists solely of the legal case set, without any cited legal provisions. It is employed with both LLMs (Direct) and LRMs.

D Supplementary Experimental Results

D.1 Set Size Sensitivity (Supplement)

This section presents additional data on set size sensitivity for other inductive pipelines: CoT (Figure 10), Long-CoT (Figure 11), and SILVER (Figure 12). These results further support the conclusions drawn in Section 5.3.

D.2 Case Study

To illustrate the structure and quality of cases within our dataset, we present representative examples of both a criminal case (Figure 14) and a civil case (Figure 15). Each case is processed using the full case-processing pipeline introduced in Section 4.3.

D.3 Error Analysis

To further assess the effect of fine-tuning on legal rule induction, we perform a comparative analysis of inference outputs from the `Llama-3.2-3B` model before and after training on the LRI-AUTO dataset. The comparison, shown in Figure 13, serves as a focused *error analysis* illustrating how fine-tuning reduces hallucinations, aligns induced rules with ground-truth logic, and improves the accuracy of legal terminology and structural coherence. Before fine-tuning, the `Llama-3.2-3B` model exhibits several systematic deficiencies:

- **Semantic Drift:** The model frequently hallucinates domain-irrelevant concepts (e.g., tax obligations or traffic negligence) that are absent from the case facts. This reflects an inability to constrain generation to the domain-specific legal frame.
- **Structure Inconsistency:** Generated rules often deviate from the tripartite logical form (*Hypothetical Condition–Behavior Pattern–Legal Consequence*), merging components or omitting legal outcomes.
- **Terminological Hallucination:** The pretrained model uses incorrect or generic legal terms (e.g., "increase responsibility" rather than "sentence aggravation"), diminishing alignment with professional terminology.

After fine-tuning on the LRI-AUTO dataset, these issues are largely mitigated:

- Generated rule structures adhere closely to the expected tripartite form.
- Hallucinations and off-topic insertions drop sharply. Induced rules align with actual case fact patterns.
- Implicit rules (e.g., voluntary surrender, compensation) emerge with appropriate conditional logic and proportionate legal consequences.
- Both procedural and substantive law distinctions are captured, with consistent and context-sensitive vocabulary.

The comparative analysis (Figure 13) evidences a substantial shift from generic, loosely structured outputs to logically coherent, legally valid formulations. This improvement supports the hypothesis that large-scale LRI datasets can significantly enhance inductive capacity in LLMs.

Judge Quality Assessment Question	Yes %
Is the assessment of element completeness correct?	100.0%
Is the assessment of sensitive content correct?	100.0%
Is the assessment of rule coverage correct?	98.24%
Is the final assessment conclusion correct?	97.36%

Table 6: Human analysis of DeepSeek-V3 judge quality.

Rule Quality Review Question	Yes (%)
Is the explicit rule applicable to all the cases in its set?	94.21%
Is the implicit rule applicable to more than half of the cases in its set?	95.03%
Is the rule logically consistent and does it use legal terminology appropriately?	99.59%
Is the rule distinct and not redundant with other rules?	100.0%
Are all fields in this rule correct?	93.63%

Table 7: Human evaluation of LRI-AUTO data quality.

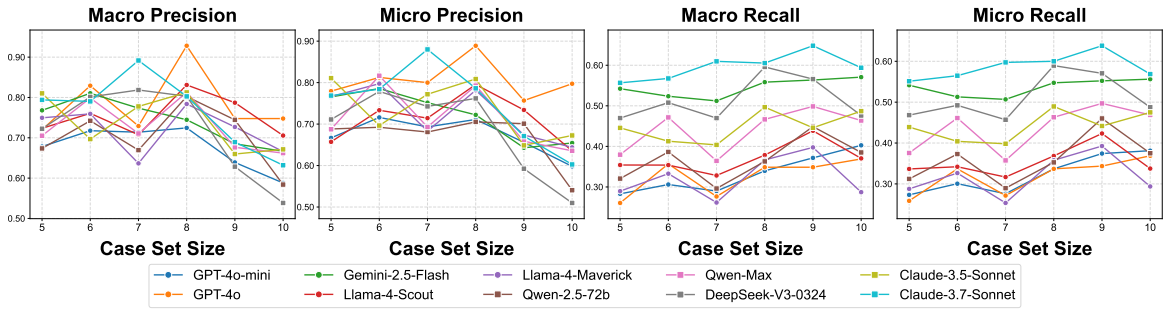

Figure 10: Performance trends of CoT of ten LLMs across varying case set sizes.

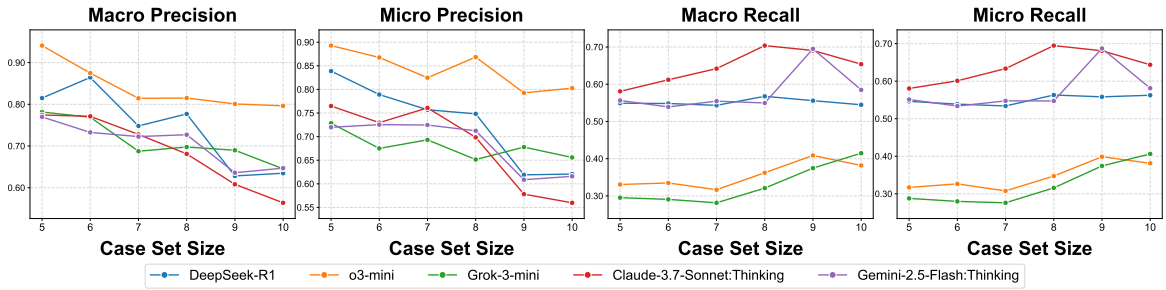

Figure 11: Performance trends of Long-CoT of five LRMs across varying case set sizes.

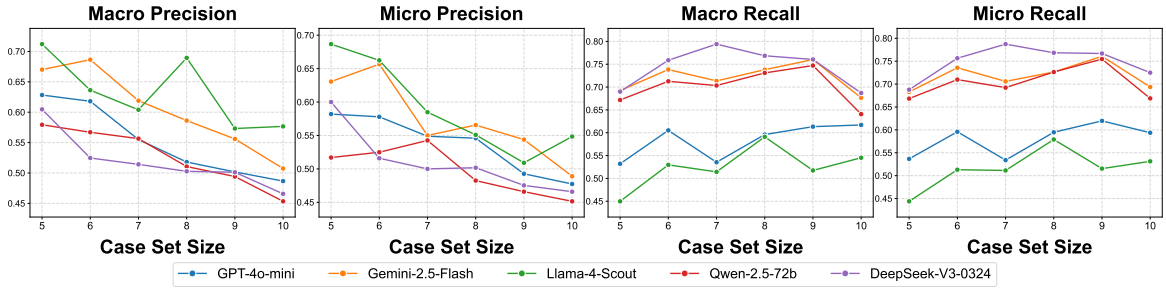

Figure 12: Performance trends of SILVER of five LLMs across varying case set sizes.

A legal case typically includes a description of the facts, legal analysis, relevant legal provisions, and the ruling. Please read the given legal case and extract the following four parts: Fact Description, Litigation Process, Legal Analysis, and Judgment Result.
[Element Definitions]
Fact Description: The basic circumstances of the case and the core dispute (maintaining the integrity of the events).
Litigation Process: The trial process and procedural matters of the case.
Legal Analysis: The reasoning process of the judgment (reflecting the logic of legal application).
Judgment Result: The final disposition and conclusion.

Please process the following case:
{Legal Case}
The legal provisions cited in this case are as follows:
{Legal Provisions}

[Extraction Rules]
(1) Content Requirements

- Prioritize using the original wording; key details must not be omitted. Do not summarize; do not summarize; do not summarize.
- The legal analysis must reflect the logic of how the provisions were applied, but specific article numbers/content of the provisions should not appear.
- Direct citation of charges or legal terms is prohibited (e.g., use "caused property loss to others" instead of "theft"). This is especially true for the Judgment Result section.
- There should be no redundant information or logical contradictions among the four parts.
- The four parts should be able to corroborate each other and the cited legal provisions, reflecting the application logic of the legal provisions.

(2) Desensitization Norms

- Replace all entities with pseudonyms (People: A/B/C; Organizations: Company A/Unit B; Locations: Place C).
- Basic identity information such as gender, age, and occupation should be retained.
- Remove court information (replace specific court names with "adjudicating authority"); remove personal information of judges, lawyers, etc.

(3) Output Format
Output according to the following JSON format:

```
{
    "Fact Description": "XXX",
    "Litigation Process": "XXX",
    "Legal Analysis": "XXX",
    "Judgment Result": "XXX"
}
```

Table 8: Prompt of case content structuring.

Please extract legal rules from the following set of legal cases and the corresponding legal provisions, and output in the required format.

[Element Definitions]

Each legal rule must contain the following three components:

1. Hypothetical Conditions: Conditions and circumstances under which the rule applies, including applicable subjects and their behaviors.

2. Behavioral Pattern: Specifies how people should act, including permissive, obligatory, and prohibitive patterns.

- **Permissive pattern:** Uses expressions such as "may," "is entitled to," or "is allowed to."
- **Obligatory pattern:** Uses expressions such as "shall," "must," or "has the obligation to."
- **Prohibitive pattern:** Uses expressions such as "prohibited," "shall not," or "must not."

3. Legal Consequence: Specifies the consequences of complying or not complying with the behavioral pattern.

- Positive consequence: Legal effect resulting from compliance.
- Negative consequence: Legal liability resulting from violation.

Here are examples of the three behavioral patterns:

1. **Permissive:**

Hypothetical Condition: A natural person wishes to engage in a civil transaction.
Behavioral Pattern: The person may (but is not required to) enter into a contract.
Legal Consequence: If a contract is formed, the person is bound by it; if not, there is no contractual obligation.

2. **Obligatory:**

Hypothetical Condition: Citizens, legal persons, or other organizations meet the conditions for tax liability (e.g., taxable income).
Behavioral Pattern: Must pay taxes on time and in full.
Legal Consequence: If taxes are paid lawfully, rights are enjoyed normally; if not, there may be fines, late fees, or other liabilities.

3. **Prohibitive:**

Hypothetical Condition: A natural person with full criminal responsibility.
Behavioral Pattern: Prohibited from committing theft.
Legal Consequence: If no theft is committed, there is no liability; if theft occurs, the person may face criminal penalties, such as detention, fines, or imprisonment.

[Extraction Rules]

I. Explicit Rule Extraction

- Must directly correspond to the cited legal provisions and reflect their core content;
- May combine multiple relevant provisions into a composite rule;
- Direct reference to specific article numbers or content is prohibited; instead, summarize into a general rule applicable to the case set;
- Explicit rules must apply to all cases in the set.

II. Implicit Rule Extraction

- Must be inferred from commonalities among cases and not directly derived from legal provisions;
- Should reflect discretionary standards in judicial practice;
- Must apply to most cases in the set (i.e., more than half).

Example: From all traffic accident cases, infer that "if the driver fails to exercise reasonable care, liability may be increased."

III. General Requirements

- Each rule must include all three components to form a complete logical chain: "If [Hypothetical condition], then [behavioral pattern], and [legal consequence] follows."
- Type must be one of: Criminal / Civil / Procedural; do not use other types.
- Avoid duplication; merge similar rules.
- Do not omit rules, especially those clearly reflected in the cited legal provisions.
- Do not use legal terminology or charges directly (e.g., use "caused property loss to others" instead of "theft").

Table 9: Prompt of legal rule extraction from case set.

The following is the set of legal cases:
{Legal Case Set}
The legal provisions cited in the case set are as follows:
{Legal Provisions}

[Output Format]
Please output in the following JSON format:

```
{
  "Explicit Rules": [
    {
      "Applicable Case Count": 10,
      "Type": "Criminal",
      "Content": {
        "Hypothetical Condition": "A natural person with full criminal responsibility",
        "Behavioral Pattern": {
          "Type": "Prohibitive",
       "Description": "Prohibited from intentionally and unlawfully depriving others of life"
        },
     "Legal Consequence": "If a person kills, they may face the death penalty, life imprisonment,
        or fixed-term imprisonment of over ten years"
      }
    }
  ],
  "Implicit Rules": [
    {
      "Applicable Case Count": 10,
      "Type": "Criminal",
      "Content": {
        "Hypothetical Condition": "The suspect has voluntarily surrendered",
        "Behavioral Pattern": {
          "Type": "Obligatory",
          "Description": "Should truthfully confess the main facts of the offense"
        },
       "Legal Consequence": "May receive a lighter or mitigated punishment according to law"
      }
    }
  ],
  "Unreflected Provisions": {
    "Civil Code of the People's Republic of China": ["Article 111"],
    "Civil Procedure Law of the People's Republic of China": ["Article 120", "Article 131"]
  }
}
```

Table 10: Prompt of legal rule extraction from case set. (Continue)

Please extract legal rules from the following set of legal cases and output in the required format.
[Element Definitions]
Each legal rule must contain the following three components:

- **1. Hypothetical Conditions:** The part of a legal rule concerning the conditions and circumstances for its application, including conditions for application and conditions for the subject's behavior.
- **2. Behavioral Pattern:** The part of a legal rule that specifies how people should act, including permissive (authorization) patterns, obligatory (duty) patterns, and prohibitive (prohibition) patterns.

  - **Permissive pattern:** Uses authorizing expressions such as "may," "is entitled to," or "is allowed to."

  - **Obligatory pattern:** Uses mandatory expressions such as "shall," "must," or "has the obligation to."

  - **Prohibitive pattern:** Uses prohibitive expressions such as "prohibited," "shall not," or "must not."
- **3. Legal Consequence:** The part of a legal rule that specifies the corresponding results people should bear when their actions comply with or violate the requirements of the behavioral pattern.

  - **Positive consequence:** The legal effect resulting from compliance with the behavioral pattern.

  - **Negative consequence:** The legal liability resulting from violation of the behavioral pattern.

Here are examples of legal rules for the three behavioral patterns:

- **1. Permissive:**

  Hypothetical Condition: A natural person wishes to engage in a civil transaction.

  Behavioral Pattern: The natural person may (but is not required to) enter into a contract.

  Legal Consequence: If a contract is entered into, they are legally bound by the contract; if no contract is entered into, there is no contractual obligation.
- **2. Obligatory:**

  Hypothetical Condition: Citizens, legal persons, and other organizations meet the conditions for tax liability (e.g., have taxable income).

  Behavioral Pattern: Must pay taxes on time and in full.

  Legal Consequence: If taxes are paid according to law, rights are enjoyed normally; if taxes are not paid according to law, they may face fines, late fees, or other legal liabilities.
- **3. Prohibitive:**

  Hypothetical Condition: A natural person with full criminal responsibility.

  Behavioral Pattern: Prohibited from committing theft.

  Legal Consequence: If no theft is committed, there is no legal liability; if theft is committed, they may face criminal penalties, such as detention, fines, or fixed-term imprisonment.

[Extraction Rules]
1. Each rule must include all three components, forming a complete logical chain:
"If [Hypothetical condition], then [behavioral pattern], then/otherwise [legal consequence]."
2. Do not use specific article numbers, content, or charges; summarize into a general rule applicable to the given case set.
3. Must be inferred from commonalities among cases and should reflect discretionary standards in judicial practice.
4. The extracted rules must apply to ≥ 51% of the cases.
Example: Infer from all traffic accident cases in the set that "if the driver fails to exercise reasonable care, liability may be increased."
5. Combining multiple relevant provisions to form a composite rule is allowed.
6. Type annotation: Criminal / Civil / Procedural; do not use other types.
7. Avoid duplication; merge similar rules.

Table 11: Prompt of legal rule induction from a case set in the evaluation phase.

The set of legal cases is as follows:
{Legal Case Set}

[Output Format]
Please output in the following JSON format:

```
{
  "Extracted Rules": [
    {
      "Type": "Criminal",
      "Content": {
        "Hypothetical Condition": "A natural person with full criminal responsibility",
        "Behavioral Pattern": {
          "Type": "Prohibitive",
       "Description": "Prohibited from intentionally and unlawfully depriving others of life"
        },
      "Legal Consequence": "If a person kills, they face the death penalty, life imprisonment,
        or fixed-term imprisonment of over ten years"
      }
    },
    {
      "Type": "Procedural",
      "Content": {
        "Hypothetical Condition": "The plaintiff in a civil case files a lawsuit",
        "Behavioral Pattern": {
          "Type": "Obligatory",
       "Description": "Shall provide clear claims and factual reasons when filing the lawsuit"
        },
        "Legal Consequence": "If the requirements are met, the case shall be accepted;
        if the requirements are not met, a one-time notice for correction shall be given"
      }
    }
  ]
}
```

Table 12: Prompt for legal rule induction from case set in the evaluation phase (Continue).

Please extract legal rules from the following set of legal cases and output in the required format.
[Element Definitions]
Each legal rule must contain the following three components:

- **1. Hypothetical Conditions:** The part of a legal rule concerning the conditions and circumstances for its application, including conditions for application and conditions for the subject's behavior.
- **2. Behavioral Pattern:** The part of a legal rule that specifies how people should act, including permissive (authorization) patterns, obligatory (duty) patterns, and prohibitive (prohibition) patterns.

  - **Permissive pattern:** Uses authorizing expressions such as "may," "is entitled to," or "is allowed to."

  - **Obligatory pattern:** Uses mandatory expressions such as "shall," "must," or "has the obligation to."

  - **Prohibitive pattern:** Uses prohibitive expressions such as "prohibited," "shall not," or "must not."
- **3. Legal Consequence:** The part of a legal rule that specifies the corresponding results people should bear when their actions comply with or violate the requirements of the behavioral pattern.

  - **Positive consequence:** The legal effect resulting from compliance with the behavioral pattern.

  - **Negative consequence:** The legal liability resulting from violation of the behavioral pattern.

Here are examples of legal rules for the three behavioral patterns:

- **1. Permissive:**

  Hypothetical Condition: A natural person wishes to engage in a civil transaction.

  Behavioral Pattern: The natural person may (but is not required to) enter into a contract.

  Legal Consequence: If a contract is entered into, they are legally bound by the contract; if no contract is entered into, there is no contractual obligation.
- **2. Obligatory:**

  Hypothetical Condition: Citizens, legal persons, and other organizations meet the conditions for tax liability (e.g., have taxable income).

  Behavioral Pattern: Must pay taxes on time and in full.

  Legal Consequence: If taxes are paid according to law, rights are enjoyed normally; if taxes are not paid according to law, they may face fines, late fees, or other legal liabilities.
- **3. Prohibitive:**

  Hypothetical Condition: A natural person with full criminal responsibility.

  Behavioral Pattern: Prohibited from committing theft.

  Legal Consequence: If no theft is committed, there is no legal liability; if theft is committed, they may face criminal penalties, such as detention, fines, or fixed-term imprisonment.

[Extraction Rules]
1. Each rule must include all three components, forming a complete logical chain:
"If [Hypothetical condition], and [behavioral pattern], then/otherwise [legal consequence]."
2. Do not use specific article numbers, content, or charges; summarize into a general rule applicable to the given case set.
3. Must be inferred from commonalities among cases and should reflect discretionary standards in judicial practice.
4. The extracted rules must apply to ≥51% of the cases.
Example: Infer from all traffic accident cases in the set that "if the driver fails to exercise reasonable care, liability may be increased."
5. Combining multiple relevant provisions to form a composite rule is allowed.
6. Type annotation: Criminal / Civil / Procedural; do not use other types.
7. Avoid duplication; merge similar rules.

Table 13: Prompt of new rule induction.

The set of legal cases is as follows:
{Legal Case Set}
The rules already extracted are as follows, please do not extract them again:
{Already Extracted Rules}
Please do not extract existing rules again to avoid redundancy.

[Output Format]
Please output in the following JSON format:

```
{
  "Extracted Rules": [
    {
      "Type": "Criminal",
      "Content": {
        "Hypothetical Condition": "A natural person with full criminal responsibility",
        "Behavioral Pattern": {
          "Type": "Prohibitive",
       "Description": "Prohibited from intentionally and unlawfully depriving others of life"
        },
      "Legal Consequence": "If a person kills, they face the death penalty, life imprisonment,
        or fixed-term imprisonment of over ten years"
      }
    },
    {
      "Type": "Procedural",
      "Content": {
        "Hypothetical Condition": "The plaintiff in a civil case files a lawsuit",
        "Behavioral Pattern": {
          "Type": "Obligatory",
       "Description": "Shall provide clear claims and factual reasons when filing the lawsuit"
        },
        "Legal Consequence": "If the requirements are met, the case shall be accepted;
        if the requirements are not met, a one-time notice for correction shall be given"
      }
    }
  ]
}
```

Table 14: Prompt of new rule induction (Continue).

Please verify the applicable case count and structural integrity of the following legal rules based on the given set of legal cases and legal rules.
[Element Definitions]
Each legal rule must contain the following three components:

- **1. Hypothetical Conditions:** The part of a legal rule concerning the conditions and circumstances for its application, including conditions for application and conditions for the subject's behavior.
- **2. Behavioral Pattern:** The part of a legal rule that specifies how people should act, including permissive (authorization) patterns, obligatory (duty) patterns, and prohibitive (prohibition) patterns.
  - **Permissive pattern:** Uses authorizing expressions such as "may," "is entitled to," or "is allowed to."
  - **Obligatory pattern:** Uses mandatory expressions such as "shall," "must," or "has the obligation to."
  - **Prohibitive pattern:** Uses prohibitive expressions such as "prohibited," "shall not," or "must not."
- **3. Legal Consequence:** The part of a legal rule that specifies the corresponding results people should bear when their actions comply with or violate the requirements of the behavioral pattern.
  - **Positive consequence:** The legal effect resulting from compliance with the behavioral pattern.
  - **Negative consequence:** The legal liability resulting from violation of the behavioral pattern.

The set of legal cases is as follows:
{Legal Case Set}
The rule set to be evaluated is as follows:
{Rule Set to be Evaluated}

Output only the JSON-formatted content; do not add any explanatory text.
[Output Format]

```
{
  "Evaluation Results": [
    { "Rule ID": 1, "Applicable Case Count": 10 (Assumed value, should be calculated
    based on the case set), "Rule Integrity": "Complete"/"Incomplete"},
    { "Rule ID": 2, "Applicable Case Count": 7 (Assumed value, should be calculated
    based on the case set), "Rule Integrity": "Complete"/"Incomplete"},
    { "Rule ID": 3, "Applicable Case Count": ...

  ]
}
```

Table 15: Prompt for legal rule verification.

Multi-dimensional assessment of target rules based on a legal rule quality assessment framework.

[Assessment Object]
Rule to be assessed:
`{Rule to be assessed}`
Reference Rule Sets:
Explicit Rule Set (directly corresponding to legal articles):
`{Explicit rule set}`
Implicit Rule Set (judicial practice conventions):
`{Implicit rule set}`

[Assessment Criteria]
1. Three-element check

- **Hypothetical Condition:** Whether the preconditions for rule application are clearly defined.
- **Behavioral Pattern:** Whether the type (may do/should do/must not do) is accurately marked and described.
- **Legal consequences:** Whether it includes the positive and negative consequences corresponding to the Behavioral Pattern.

2. Prohibited content check

- Whether there are prohibited references such as legal article numbers, names of crimes, etc.

3. Rule coverage check

- Whether it is logically equivalent to any rule in the explicit rule set.
- Whether it is logically equivalent to any rule in the implicit rule set.

4. Assessment conclusion

- Rules that meet all the above requirements are **"Correct"**.
- In the coverage check, **"logical equivalence"** must be achieved to be considered **"Correct"**.
- If it does not meet the three-element check or contains prohibited content, it is **"Incorrect"**.
- If it does not match any explicit or implicit rules, it is **"Incorrect"**.

[Output Format]

```
{
  "Element Completeness": {
    "Hypothetical Condition": "Not Present"/"Correct"/"Incorrect",
    "Behavioral Pattern": "Not Present"/"Correct"/"Incorrect",
    "Legal Consequences": "Not Present"/"Correct"/"Incorrect"
  },
  "Prohibited/Sensitive Content": "Present"/"Not Present",
  "Rule Coverage": {
    "Explicit Rules": "Logically Equivalent"/"Partially Matches"/"Does Not Match",
    "Implicit Rules": "Logically Equivalent"/"Partially Matches"/"Does Not Match"
  },
  "Assessment Conclusion": "Correct"/"Incorrect"
}
```

Table 16: Prompt of multi-dimensional assessment of legal rules.

§ Rule Set Comparison (Ground-Truth vs Induced)

§ Ground-Truth Rule Set:

- **Explicit Rule 1 (Criminal)** *Hypothetical Condition:* A natural person with full criminal responsibility commits an act endangering public safety. *Behavior Pattern:* Prohibition of endangering public safety by arson, explosion, or other dangerous methods. *Legal Consequence:* 3–10 years imprisonment; leniency for voluntary surrender, guilty plea, etc.
- **Explicit Rule 2 (Litigation Procedure)** *Hypothetical Condition:* Suspect pleads guilty and accepts punishment, no procedural flaws. *Behavior Pattern:* Charges and sentencing recommendations should be adopted. *Legal Consequence:* Court adopts recommendations unless manifestly improper.
- **Implicit Rule 1 (Criminal)** *Hypothetical Condition:* Voluntary surrender and truthful confession. *Behavior Pattern:* Should cooperate with judicial proceedings. *Legal Consequence:* Leniency (20–40% sentence reduction).
- **Implicit Rule 2 (Criminal)** *Hypothetical Condition:* Compensation for losses after endangering public safety. *Behavior Pattern:* Should obtain victim's forgiveness. *Legal Consequence:* Discretionary leniency (10–30% sentence reduction).
- **Implicit Rule 3 (Criminal)** *Hypothetical Condition:* Multiple overlapping dangers. *Behavior Pattern:* Prohibition of dangerous acts in flammable areas. *Legal Consequence:* Base sentence increased to mid-to-upper range.

§ Induced Rule Set Before Training:

- **Rule 1 (Explicit/Criminal)** *Hypothetical Condition:* Driver failed to exercise reasonable care. *Behavior Pattern:* Prohibition of inattentive driving. *Legal Consequence:* May increase responsibility.
- **Rule 2 (Explicit/Criminal)** *Hypothetical Condition:* Tax obligations apply. *Behavior Pattern:* Must pay taxes in full and on time. *Legal Consequence:* Legal rights if paid; fines or liabilities if not.
- **Rule 3 (Explicit/Criminal)** *Hypothetical Condition:* Natural person with criminal responsibility. *Behavior Pattern:* Prohibition of theft. *Legal Consequence:* Punishment if theft committed.

§ Induced Rule Set After Training:

- **Rule 1 (Explicit/Criminal)** *Hypothetical Condition:* Criminal act endangering public safety. *Behavior Pattern:* Prohibition of arson or similar acts. *Legal Consequence:* Up to 3 years imprisonment; 3–10 years for serious cases.
- **Rule 2 (Explicit/Litigation Procedure)** *Hypothetical Condition:* Guilty plea and clear facts. *Behavior Pattern:* Sentencing recommendations should be adopted. *Legal Consequence:* Judgment consistent with recommendations.
- **Rule 3 (Implicit/Criminal)** *Hypothetical Condition:* Voluntary surrender or confession. *Behavior Pattern:* Truthfully confess main facts. *Legal Consequence:* Leniency or mitigation possible.
- **Rule 4 (Implicit/Criminal)** *Hypothetical Condition:* Compensation and forgiveness. *Behavior Pattern:* Fulfill civil compensation obligations. *Legal Consequence:* Discretionary leniency.
- **Rule 5 (Explicit/Criminal)** *Hypothetical Condition:* Surrender or guilty plea. *Behavior Pattern:* Sign plea agreement and accept sentencing. *Legal Consequence:* Lenient procedures applied.

Figure 13: An example of a rule set generated by **Llama-3.2-3B** before and after training on the LRI-AUTO dataset.

§ A Criminal Case Example

Fact Description: The defendant (male, born in 1978, with a junior high school education, and working as a crew member), due to a personal dispute with Person A, set fire to dry straw in the bedroom of the house he shared with Person A at around 11:00 p.m. on February 18, 2020, after consuming alcohol. He also recorded a video of the act and sent it to Person A. The house is located in Area C and was rented by Person A. It is adjacent to Person B's residence on the west side, 1.3 meters from Person C's residence on the east, and across the street from Person D's house to the south. There was a haystack beside the street.
Litigation Process: The case was prosecuted by the public prosecution authority and publicly tried by a lawfully formed collegial panel. The prosecution alleged that the defendant's actions constituted a crime of endangering public safety, presenting evidence such as victim statements, witness testimonies, and on-site inspection records. The defendant and his defense counsel did not dispute the charges. The defense argued for leniency based on voluntary surrender and admission of guilt. The trial court confirmed eight categories of evidence presented and challenged during the hearing.
Legal Analysis: The court determined that the defendant intentionally committed arson by setting fire to another person's property, which posed a substantial danger to public safety. Although the act did not result in severe consequences, the fire occurred in a densely populated area with flammable materials nearby, presenting a real risk. The defendant voluntarily turned himself in and truthfully confessed, which constitutes a legal ground for leniency. He also voluntarily admitted guilt and accepted punishment, qualifying for a more lenient sentence. The sentencing recommendation by the prosecution was deemed appropriate given the facts and circumstances and was adopted by the court.
Judgment Result: The defendant was sentenced to three years and six months of fixed-term imprisonment, with the sentence commencing on February 19, 2020. The court considered mitigating factors such as voluntary surrender, truthful confession, and admission of guilt when determining the sentence. The time already spent in detention was credited toward the prison term.

Figure 14: A criminal case from CJO after case processing.

§ A Civil Case Example

Fact Description: On March 18, 2019, Party A (male, born August 11, 1969, Han ethnicity) applied for a loan through his electronic banking account with Bank A, signing the "Quick e-Loan Agreement" and the "Loan Service Agreement" electronically (via data message). The contract stipulated a loan amount of 71,500 RMB, with a term from March 18, 2019, to March 18, 2020, and an annual interest rate of 5.6%. In case of overdue payments, the penalty interest rate would increase by 50%. Bank A disbursed the loan as agreed, but Party A failed to make repayments according to the contract.
Litigation Process: The case was filed on April 15, 2021. The court applied summary procedures and held a public hearing on May 25, 2021. Bank A's authorized litigation representative attended the trial. Party A, though legally summoned, did not appear in court, so the court conducted a trial in absentia.
Legal Analysis: The loan agreements signed electronically by both parties reflected their true intent and contained legally valid content, making the contracts legally binding and effective. Since Bank A fulfilled its obligation by disbursing the loan, and Party A breached the agreement by failing to repay, he is liable to return the principal and pay the agreed interest and penalty interest. As for Bank A's claims for announcement and asset preservation fees, the court did not support them due to a lack of evidence proving that those expenses are actually incurred.
Judgment Result: Party A is ordered to repay Bank A the loan principal of 71,500 RMB within ten days after the judgment takes effect, along with interest and penalty interest as stipulated in the contract. Other claims made by Bank A are dismissed. If Party A fails to fulfill the monetary obligations on time, it must pay double interest on the overdue amount during the delay period. The case acceptance fee of 790 RMB is to be borne by Party A.

Figure 15: A civil case from CJO after case processing.