When Do LLMs Apply the Wrong Law? Diagnosing LLM Failures in Temporal Legal Reasoning

Yiqian Huang[1], Shuyuan Zheng[2]*, Qianying Liu[3], Shaowen Peng[4], Yuntao Kong[5], Kotaro Funakoshi[1], Chuan Xiao[2], Manabu Okumura[1], Yang Cao[1]
[1]Institute of Science Tokyo [2]Osaka University [3]NII LLMC
[4]Nara Institute of Science and Technology [5]Center of Juris-Informatics, ROIS-DS
h1k@lr.first.iir.isct.ac.jp, zheng@ist.osaka-u.ac.jp

Abstract

Legal reasoning tasks such as legal judgment prediction (LJP) require identifying the temporally correct version of the law governing a case—a capability we term *temporal applicable-law determination*. However, whether large language models (LLMs) can reliably perform this task remains unexplored. In this paper, we construct a benchmark to evaluate LLMs on temporal applicable-law determination, and systematically investigate why they fail at temporal legal reasoning. Our experiments reveal four key findings. First, LLMs exhibit a strong bias toward applying the most recently enacted law, regardless of when the legally relevant facts occurred. Second, this bias does not stem from an inability to understand that laws have temporal scope, nor from a lack of knowledge about historical statutes. Third, we provide behavioral evidence that reinforcement-learning-shaped explicit reasoning may be a key mechanism: while improving general reasoning ability, it reduces the diversity of reasoning paths, causing models to converge on applying the current law. Fourth, this produces a counterintuitive inverse relationship: models with stronger general reasoning ability tend to perform *worse* on temporal legal reasoning. Our findings offer concrete guidance for future work on improving LLM performance in temporally grounded legal reasoning[1].

1 Introduction

The rapid advancement of large language models (LLMs) has spurred growing interest in automating legal reasoning, with significant implications for judicial efficiency and legal accessibility (Cui et al., 2023; Huang et al., 2023). Researchers have explored LLM-based approaches across a wide spectrum of legal tasks, including legal judgment prediction (LJP) (Luo et al., 2017), legal fact prediction (Liu et al., 2025), court opinion generation (Li and Zhang, 2021), and contract analysis (Hendrycks et al., 2021). These efforts have demonstrated that LLMs can serve as powerful tools in the legal domain, motivating broader deployment in legal practice.

A critical yet underexplored dimension of legal reasoning is its inherently *temporal* nature. The governing principle of *non-retroactivity* dictates that legal events should generally be adjudicated under the law in force at the time the relevant conduct occurred (Bowen, 2005). Even when a statute has since been revised, its earlier version may remain operative for events predating the revision. Misapplying the wrong temporal version is not a clerical error: it can fundamentally alter the legal characterization of conduct and ultimately the outcome of a case. Accurate legal reasoning therefore requires models to identify which statutory version governs a given set of facts—a capability we term *temporal legal reasoning*.

Despite its importance, only a handful of existing studies have considered the time dimension in LLM-based legal analysis (Barale et al., 2025; Santosh and Vuong, 2025; Han et al., 2025). Among them, in LawShift, Han et al. (2025) found that even state-of-the-art (SOTA) LLMs consistently fail to incorporate synthetic statutory amendments into their legal judgments, defaulting instead to predictions anchored in the current law. However, they stop short of diagnosing *why* these failures occur, leaving the root cause unexplained.

This paper investigates why LLMs fail to identify and apply the temporally correct version of the law. Identifying the temporally applicable law is a necessary prerequisite for downstream legal reasoning tasks such as LFP and LJP: without first establishing which statutory version governs a case, any subsequent judgment prediction or legal analysis is built on an unreliable foundation. To this end, we construct a benchmark for *temporal applicable-*

*Corresponding author.
[1]Our code and data will be made available upon acceptance.

law determination (TALD)—requiring models to identify, given a case's facts and the date of the relevant events, which statutory version governs the case. Our empirical investigation yields four key findings. **First**, LLMs exhibit a strong and systematic bias toward applying the most recently enacted law, regardless of when the legally relevant facts occurred—directly explaining the failures observed in LawShift. **Second**, this bias does not stem from ignorance of prior statutory versions, nor from a failure to understand that laws have temporal scope. **Third**, we identify reinforcement learning (RL) as a key mechanism: while it improves general reasoning ability, it simultaneously reduces the diversity of reasoning paths, causing models to converge almost exclusively on applying the current law. **Fourth**, this produces a counterintuitive result: models with stronger general reasoning ability tend to perform *worse* on temporal legal reasoning.

The contributions of this paper are as follows. **(1)** To the best of our knowledge, this is the first work to systematically diagnose the root causes of LLM failures in temporal legal reasoning. **(2)** We introduce a benchmark for temporal applicable-law determination, enabling controlled evaluation of whether LLMs correctly identify the temporally appropriate statutory version for a given case. **(3)** Our findings suggest that reasoning-oriented RL can reduce diversity in TALD-relevant reasoning paths, causing models to converge toward the current-law trajectory.

2 The Temporal Applicable-Law Determination Task

2.1 Preliminary

Time is a fundamental dimension of legal reasoning. Statutes and regulations are enacted, amended, and repealed at specific points in time, and their legal force is bounded by corresponding effective periods. A foundational principle governing this temporal structure is *non-retroactivity* (Kryvoi and Matos, 2021): newly enacted law generally applies only to facts and legal relationships arising after its entry into force, not to those already completed under prior law. Consequently, determining *which version* of a law governs a given dispute is a prerequisite for any legally sound conclusion.

Temporal applicability determination is a foundational prerequisite for virtually all legal reasoning. Before a model can apply a statute to reach a legal conclusion, it must first identify the version of that statute that was in effect at the legally relevant time. Applying an incorrect version—whether superseded or not yet in force—leads to legally erroneous results regardless of the quality of the substantive reasoning. Temporal applicable-law determination is therefore a necessary first step in the legal reasoning pipeline.

2.2 Task Definition

Consider that we have a collection of statutes $N = \{1, \ldots, n\}$, where each statute $i \in N$ may have multiple versions due to temporal legal evolution. Given a query (F, Q), where F denotes the facts of a case and Q denotes the legal question by the user, the *Temporally Applicable-Law Determination (TALD)* task aims to identify a sequence of legal citations $\boldsymbol{y} = [y_1, \ldots, y_n]$ where y_i denotes the correct version of statue i that is legally applicable to answering Q based on F. Note that if $y_i = -1$, it means that statute i is not cited in the legal reasoning process.

Let $\hat{\boldsymbol{y}} = \{\hat{y}_1, \ldots, \hat{y}_n\}$ denote the sequence of statute versions predicted by the model in the TALD task. The model performance on TALD can be measured by the TALD accuracy, defined by the matching accuracy between the ground truth $\boldsymbol{y}$ and the prediction $\hat{\boldsymbol{y}}$:

$$ACC(\hat{\boldsymbol{y}}; \boldsymbol{y}) = \frac{\sum_{i \in N} \mathbf{1}(y_i = \hat{y}_i \neq -1)}{|\{i \mid y_i \neq -1 \text{ or } \hat{y}_i \neq -1\}|}. \quad (1)$$

The objective of TALD is therefore to maximize the TALD accuracy.

3 Analysis Setup

3.1 Research Questions

We investigate the capability of large language models (LLMs) on the Temporal Applicable-Law Determination (TALD) task through three research questions (RQs). **RQ1**: *Can LLMs correctly determine the temporally applicable version of law based on the temporal information present in case facts?* This examines whether models can leverage temporal cues in case descriptions—such as the timing of legal acts or disputes—to select the correct statutory version. **RQ2**: *Do LLM failures on TALD stem from deficiencies in knowledge memorization?* Specifically, we examine whether models lack the necessary knowledge of temporal applicability rules governing Chinese civil law. **RQ3**: *Do LLM failures on TALD stem from deficiencies in legal reasoning?* Even when models possess relevant

legal knowledge, they may fail to correctly apply it to determine the temporally applicable statute.

3.2 Dataset Construction

To evaluate LLMs on the TALD task, we collect civil judgment documents from the China Judgment Online platform.[2] For each judgment, we require the model to identify the temporally correct version of applicable civil law given the case facts and the plaintiff's claims. Since Chinese civil judgments follow a standardized structure, we use regular expression matching to automatically extract case facts and plaintiff claims as TALD queries, and extract the cited civil law version as the ground-truth label. After automatic extraction, two legal experts conducted cross-validation via random sampling to verify extraction accuracy.

The resulting dataset contains 26,000 civil judgments, balanced between two splits: the *post-Code split*, comprising cases governed by the *Civil Code of the PRC* enacted in 2021, and the *pre-Code split*, comprising cases governed by prior standalone civil statutes such as the *Property Law* and *Contract Law*. Since the Civil Code essentially consolidates these prior statutes—its property and contract books correspond directly to the former *Property Law* and *Contract Law*—we treat the Civil Code as a collection of multiple statutes, each aligned with a prior standalone statute, enabling more fine-grained analysis of model performance on TALD.

To address RQ2, two legal experts additionally construct 16 either-or questions covering the legal provisions on temporal applicability of the Civil Code under Chinese law. These questions assess whether LLMs possess the requisite knowledge of temporal applicability rules, providing a diagnostic lens on the source of model failures.

3.3 Experimental Configurations

Models. We evaluate a broad range of state-of-the-art reasoning models. For proprietary models, we include GPT-5.4, Claude Opus 4.6, Gemini-3.1-Pro, DeepSeek-V3.2, and GLM-4.7. For open-source models, we evaluate Qwen3 (235B, 80B, 30B). We also include LegalOne-8B, a domain-specific legal reasoning model. By default, all models are evaluated at their highest available reasoning effort setting.

Test Data. To ensure robustness, we repeat each experiment four times and report average results.

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

Model	Mode	Old-version	New-version
		ACC	ACC
DeepSeek-V3.2	thinking	0.092	0.746
GLM-4.7	thinking	0.080	0.787
GPT-5.4	high	0.237	0.701
Gemini3.1-pro	high	0.135	0.843
Claude opus 4.6	max	0.014	0.820
Qwen3-235B	thinking	0.148	0.730
Qwen3-80B	thinking	0.028	0.798
Qwen3-30B	thinking	0.020	0.765
LegalOne-8B	thinking	0.093	0.800

Table 1: LLMs' performance of TALD on both splits, evaluated with TALD accuracy metric (Equation 1).

In each run, we randomly sample 100 cases from the post-Code split and 100 from the pre-Code split, forming a balanced test set of 200 cases.

4 Analysis on RQ1

As the exploration of Research Question I, we conduct an initial test to examine whether the ability of most advanced reasoning Large Language Models to perform TALD.

4.1 Results

4.1.1 Basic Results

Table 1 shows a sharp asymmetry between the two splits. On the Old-version split, every evaluated model scores below 0.25, and seven out of eight score below 0.15; Claude-Opus-4.6 in particular drops to 0.014. On the New-version split the same models score between 0.70 and 0.84. The gap is not explained by the difficulty of the task in general: in both splits, recovering the applicable law requires the same kind of legal reasoning over the same case facts, and the gold answer in the New-version split is simply the Civil Code book that subsumes the relevant area. The asymmetry instead indicates a directional bias—LLMs preferentially apply the newest version regardless of when the legally relevant facts occurred.

Finding 1. Advanced reasoning LLMs have directional failure in TALD. They systematically perform poorly when an older version of the applicable law should govern the case, while performing competitively when the newer version applies. This asymmetry is consistent with a strong default tendency toward citing the most recently enacted law.

4.1.2 Fault-Type Analysis

Results above show that models fail old-version cases, but do not yet reveal whether the failure is

Model	Old→New	New→Old	Non-ver.
DeepSeek-V3.2	0.85	0.03	0.12
GLM-4.7	0.89	0.00	0.11
GPT-5.4	0.69	0.14	0.17
Gemini3.1-pro	0.79	0.06	0.15
Claude-opus-4.6	0.89	0.00	0.11
Qwen3-235B	0.75	0.05	0.20
Qwen3-80B	0.87	0.00	0.13
Qwen3-30B	0.79	0.02	0.19
LegalOne-8B	0.94	0.00	0.06

Table 2: Fault type statistics. Old→New, New→Old, Non-ver respectively stand for the ratio of tested samples with OLD-TO-NEW, NEW-TO-OLD and NON-VERSIONAL type of fault.

version-centric: do LLMs show basic awareness that the case concerns temporally versioned legal sources, and focus on determining them? We therefore analyze the wrong predictions on both splits. For each missed gold target law in LLM's test sample, we classify the fault into three categories: OLD-TO-NEW, where the gold version is a previous law but the model cites a new version; NEW-TO-OLD, where the reverse occurs; and NON-VERSIONAL, where the model failed with an answer that contains only versionally unrelated laws.

Table 2 shows that only a small portion of failures are version-irrelevant, while OLD-TO-NEW type takes the majority. Across models, this category accounts for the dominant share of fault mass, while NEW-TO-OLD errors are rare. This indicates that models often identify the relevant legal family or versional counterpart, but select the wrong temporal direction. In other words, the models are not merely citing unrelated law. They display basic awareness that the case concerns a temporally versioned legal source, yet their default selection is biased toward the newer version.

Finding 2. Advanced reasoning LLMs' failures in TALD are version-centric. They often locate the relevant statute family, but choose the newer version when the previous version should apply.

5 Experiments on RQ2

RQ1 shows that LLMs fail TALD in a directional way: they tend to apply newer law versions even when previous versions should govern the case. We next examine what drives this failure. We focus on two fundamental and mostly orthogonal capabilities of reasoning LLMs as remaining explanations: whether models lack the necessary legal knowledge, and whether their reasoning capability or reasoning policy is responsible for the failure.

5.1 Legal and Temporal-Effect Knowledge

Hypothesis 1. LLMs may fail TALD because they lack the relevant legal knowledge. This may take two forms: they may not remember the content of older statutory provisions, or they may not know the temporal-effect rules that determine whether old or new law should apply.

Probe A: statutory-text memorization. We first test whether models remember the relevant statutory texts. On failed old-version cases, we ask the same model to reproduce two sets of provisions: the articles from its own wrong prediction and the gold articles under the correct previous-version law. We then compare the generated content with the official statutory text using character-level F1, ROUGE-L, and edit similarity.

Results. Table 3 shows that most models can reproduce both predicted and gold provisions with high similarity to the official text. Importantly, gold previous-version articles are not systematically less memorized than the wrongly predicted articles. For example, DeepSeek-V3.2 obtains 0.963 character-level F1 on gold articles and 0.901 on predicted articles; Qwen3-235B obtains 0.902 on gold articles and 0.776 on predicted articles. This suggests that models often possess the old-law textual knowledge needed for TALD, but still fail to select it in context.

Probe B: temporal-effect rule knowledge. Remembering statutory text is not sufficient for TALD; models must also know the legal principles governing temporal applicability. We therefore construct a 16-question multiple-choice question (MCQ) probe based on expert-designed questions about civil-law temporal-effect rules according to relevant judicial interpretation. Each question asks the model to choose the legally correct option under a specified temporal-effect condition.

Results. Table 4 shows that models generally perform well on this rule-knowledge probe. All models achieve accuracy above 0.80, and several reach around 0.90 or higher. Even models that fail substantially on old-version TALD cases can answer many abstract temporal-effect questions correctly. This indicates that TALD failure is not primarily caused by the absence of explicit knowledge about temporal-effect doctrines.

Model	Mode	C-F1(Pred.)	C-F1(Gold)	R-L(Pred.)	R-L(Gold)	Edit(Pred.)	Edit(Gold)
DeepSeek-V3.2	reasoner	0.901	0.963	0.885	0.963	0.850	0.941
GLM-4.7	thinking	0.976	0.951	0.976	0.950	0.960	0.936
GPT-5.4	high	0.876	0.834	0.875	0.831	0.806	0.754
Gemini3.1-pro	high	0.995	0.996	0.995	0.996	0.991	0.992
Claude-opus-4.6	max	1.000	0.988	1.000	0.987	1.000	0.981
Qwen3-235B	thinking	0.776	0.902	0.776	0.901	0.694	0.862
Qwen3-80B	thinking	0.976	0.984	0.974	0.983	0.955	0.970
Qwen3-30B	thinking	0.752	0.876	0.731	0.870	0.672	0.831

Table 3: Statutory knowledge memorization probe on failed previous-version cases. For each failed case, the model is asked to reproduce the content of both its predicted articles and the gold articles. C-F1, R-L, and Edit denote character-level F1, ROUGE-L, and edit similarity with the official statutory text.

Finding 3. Knowledge deficiency is not the main factor behind TALD failure. Models often possess both forms of knowledge required for TALD: the textual content of old/new statutes and abstract temporal-effect rules. Their failure, therefore, arises when applying this knowledge to concrete case facts.

Model	Mode	Accuracy
DeepSeek-V3.2	thinking	0.94
GLM-4.7	thinking	1.00
GPT-5.4	high	0.94
Gemini3.1-pro	high	0.98
Claude-opus-4.6	max	0.94
Qwen3-235B	thinking	1.00
Qwen3-80B	thinking	0.81
Qwen3-30B	thinking	0.88
LegalOne-8B	reasoning	0.88

Table 4: MCQ accuracy performance on the Civil Code Temporal Effect Judicial Interpretation.

5.2 Reasoning Capability

Having ruled out basic legal-knowledge deficiency, we next examine whether TALD failure is caused by insufficient reasoning capability. TALD requires models to connect legally relevant facts with the effective periods of candidate law versions. A natural explanation is therefore that the task is simply too difficult for current models, and that stronger general reasoning should improve performance.

5.2.1 Does stronger general reasoning help?

Hypothesis 2. A key factor behind TALD failure is insufficient general reasoning capability. If this hypothesis holds, stronger reasoning configurations should perform better, especially on old-version cases where temporal applicability must be inferred carefully. Here, we use *general reasoning ability* to refer to the reasoning competence targeted by general-purpose reasoning post-training and broad reasoning benchmarks. Then we derive the hypothesis below.

Probe. We compare model configurations that differ in expected general reasoning strength, including instruct versus thinking modes, lower versus higher reasoning effort, and different model sizes within the same family. All models are evaluated under the default prompt.

Results. The "Default" colums of Table 5 shows that stronger general reasoning does not consistently improve old-version TALD. In more than half of all LLMs, reasoning-oriented configurations perform worse than less-reasoning configurations. For example, Qwen3-80B drops from 0.080 in instruct mode to 0.028 in thinking mode. Across model families the best performances are obtained by smaller or less-reasoning configurations, e.g., Qwen3-30B-instruct (0.292), GPT-5.4-high (0.237), Qwen3-235B-thinking (0.148), rather than by the most reasoning-heavy configurations of the largest models.

This pattern contradicts the hypothesis that TALD failure is merely due to insufficient general reasoning ability. If stronger reasoning naturally solved the task, reasoning modes and higher effort should consistently improve old-version accuracy. Instead, they can reinforce the same failure direction identified in RQ1: applying newer law versions when previous versions should govern the case.

Finding 4. Stronger general reasoning does not solve TALD; engaging it more strongly often makes performance worse.

5.2.2 Examine Task-specific Reasoning Policy

The previous analysis shows that general reasoning ability do not result in better TALD performance. A plausible explanation is that reasoning LLMs possess relevant reasoning capacity, but their de-

fault reasoning behavior is not aligned with the task-specific version-applicability policy required by TALD.

Hypothesis 3. A key factor behind TALD failure is a misaligned task-specific reasoning policy. If this hypothesis holds, hints about temporal version applicability should improve TALD, especially for reasoning-oriented configurations that have enough capacity to use the guidance.

Probe. We evaluate models on the Old-Version Split under three different system prompt settings, with othe settings same as Section 5.2.1. The no-hint setting uses a default prompt. The weak hint and the strong hint respectively contains a brief reminder of legal principle for TALD and an detailed legal expertise instruction for TALD. Full prompts are provided in Appendix A.

Results. Table 5 shows that temporal-law hints substantially improve old-version TALD for many reasoning-oriented configurations. DeepSeek-V3.2 reasoner improves from 0.092 to 0.436, GLM-4.7 thinking improves from 0.080 to 0.435, Qwen3-80B thinking improves from 0.028 to 0.512, and Qwen3-30B thinking improves from 0.020 to 0.395. GPT-5.4 also improves from 0.237 to 0.457 under strong hint. At the same time, the effect is not a uniform prompt bonus. Some less-reasoning configurations that already perform relatively well under no hint show limited or even negative changes, such as Qwen3-30B instruct. This suggests that hints help primarily when the model has reasoning capacity but needs guidance toward the correct temporal-law policy.

Finding 5. TALD failure is closely tied to task-specific reasoning-policy misalignment. Domain-expertise hints can redirect reasoning-oriented models toward the correct temporal-applicability policy, whereas stronger general reasoning alone does not guarantee improvement.

6 Experiments on RQ3

RQ2 shows a counterintuitive phenomenon: stronger general reasoning ability does not naturally improve TALD, and can even drive LLMs towards more severe failure, with an unreasonably strong tendency to apply new versions of law. However, this still leaves a mechanism question: why does the reinforcement of general reasoning result in a worse TALD task-specific reasoning policy?

6.1 Hypothesis: Policy-Entropy Collapse

We interpret this phenomenon through the lens of exploration and exploitation in reasoning-oriented RL. Entropy in reasoning policy conveys the diversity of LLM's reasoning traces, serving as a useful signal of exploration in multi-step reasoning (Cheng et al., 2025; Zhang et al., 2025). Prior work observes that RL training for reasoning LLM may suffer from policy-entropy collapse, reducing exploratory behavior and making the policy increasingly deterministic (Cui et al., 2025). Recent empirical stuides shows that, in RLVR, task-irrelevant spurious rewards applied with clipping could decrease the LLM's policy entropy without contributing to its exploitation. (Chen et al., 2026).

We hypothesize that TALD is vulnerable to this mechanism. Newer statutes are often more salient in contemporary legal texts and semantically close to their predecessors. As a result, applying the newest law can become a high-probability reasoning trajectory. If reasoning-oriented RL reduces exploration over alternative temporal-applicability paths, it may strengthen this trajectory even when it is legally incorrect.

Hypothesis 5. Without hints, explicit reasoning could hurt TALD when version-related reasoning collapses toward a narrow newest-law trajectory.

If this hypothesis holds, interventions that recover TALD performance should be accompanied by higher entropy in version-related reasoning spans, reflecting reopened exploration over temporal-law alternatives.

6.2 Measuring TALD Policy Entropy

We use token-level policy entropy in the model's CoT as a behavioral proxy for reasoning exploration. Since TALD failure specifically concerns law versions, we do not measure entropy over the whole response. Instead, we focus on *version-related* reasoning spans, including law names, temporal markers, version descriptions, retroactivity, non-retroactivity, effective periods, and related legal-temporal expressions.

For each response y_i, let S_i denote the token positions belonging to version-related reasoning spans. At each token position t, we compute:

$$H_{i,t} = -\sum_{v \in \mathcal{V}} p(v \mid y_{i,<t}, x_i) \log p(v \mid y_{i,<t}, x_i), \tag{2}$$

where x_i is the case fact and $\mathcal{V}$ is the vocabulary.

Model	Mode	Default	Weak	Strong
DeepSeek-V3.2	chat	0.013	0.028	0.080
	reasoner	0.092	0.394	0.436
Claude-Opus-4.6	low	0.021	0.131	0.185
	medium	0.019	0.102	0.143
	high	0.026	0.099	0.154
	max	0.014	0.113	0.156
Gemini-3.1-Pro	low	0.146	0.501	0.439
	medium	0.131	0.503	0.489
	high	0.135	0.474	0.486
GPT-5.4	none	0.021	0.087	0.167
	low	0.082	0.444	0.456
	medium	0.196	0.472	0.486
	high	0.237	0.416	0.457
GLM-4.7	instruct	0.013	0.032	0.053
	thinking	0.080	0.337	0.435
Qwen3-235B	instruct	0.035	0.035	0.041
	thinking	0.148	0.330	0.323
Qwen3-80B	instruct	0.080	0.078	0.066
	thinking	0.028	0.386	0.512
Qwen3-30B	instruct	0.292	0.264	0.164
	thinking	0.020	0.350	0.395
LegalOne-8B	reasoning	0.093	0.280	0.298

Table 5: Effect of temporal-law hints on previous-version cases. Values are TALD accuracy.

The version-related entropy is:

$$H_i^{ver} = \frac{1}{|S_i|} \sum_{t \in S_i} H_{i,t}. \tag{3}$$

We report the average H^{ver} over previous-version cases.

6.2.1 Probe Setup

We basically adopt the same evaluation setup as Section. 5.2.2, report the TALD policy entropy H_i^{ver} together with model performance represented by sensitive-law recall TALD accuracy on the Old-Version Split of dataset.

6.2.2 Results and Analysis

Table 6 reports TALD accuracy and version-related policy entropy on the Old-Version Split. We interpret H^{ver} only within the same model family and intervention axis, since absolute entropy values are not directly comparable across tokenizers, probability calibration, and decoding implementations.

Hint perturbations increase version-related exploration when TALD improves. Across Qwen3 and LegalOne, stronger temporal-law hints generally increase both TALD accuracy and H^{ver}. Qwen3-30B improves from 0.020 accuracy and 0.315 entropy under no hint to 0.450 accuracy and 0.462 entropy under strong hint. Qwen3-80B improves from $0.050/1.148$ to $0.578/1.211$, and LegalOne-8B improves from $0.093/0.147$ to $0.298/0.173$. Qwen3-235B shows a saturated pattern: accuracy jumps from 0.155 to 0.524 under weak hint, while entropy increases from 0.238 to 0.253; stronger hint further increases entropy to 0.266 but does not further improve accuracy.

These patterns suggest that successful TALD correction is accompanied by reopening version-related reasoning paths, but larger entropy is not itself the objective. Once a model has reached a better task-specific policy region, additional uncertainty around version-related tokens may not produce further accuracy gains.

Finding 6. RQ3 provides behavioral evidence for a policy-entropy explanation of why explicit reasoning can hurt TALD. Reasoning-oriented policies may over-exploit a salient newest-law trajectory on old-version cases. Interventions that improve TALD are generally accompanied by higher entropy in version-related CoT spans, suggesting that successful TALD requires task-directed diversity over temporal-law alternatives rather than merely more explicit reasoning.

7 Related Work

7.1 Downstream Tasks Related to TALD

A large body of Legal AI work studies downstream prediction tasks from case facts, including legal judgment prediction, charge prediction, law-article prediction, and legal citation prediction. Early LJP benchmarks such as CAIL2018 (Xiao et al., 2018) formulate judgment prediction as predicting applicable law articles, charges, and penalties from fact descriptions. Recent work further improves legal judgment prediction by incorporating legal knowledge, multi-task learning, retrieval, or LLM-based citation prediction. These tasks all require models to identify legally relevant authorities before making downstream predictions (Fei et al., 2023; Han et al., 2026; Liu et al., 2025). TALD differs from these tasks by isolating a more basic prereq-

Model / Mode	Effort	No hint		Weak hint		Strong hint	
		ACC	H_i^{ver}	ACC	H_i^{ver}	ACC	H_i^{ver}
Qwen3-235B	thinking	0.155 (.025)	0.238 (.006)	0.524 (.025)	0.253 (.004)	0.523 (.032)	0.266 (.005)
Qwen3-80B	thinking	0.050 (.002)	1.148 (.026)	0.459 (.022)	1.185 (.029)	0.578 (.024)	1.211 (.025)
Qwen3-30B	thinking	0.020 (.015)	0.315 (.005)	0.371 (.015)	0.425 (.010)	0.450 (.020)	0.462 (.014)
LegalOne-8B	reasoning	0.093 (.008)	0.147 (.006)	0.280 (.032)	0.159 (.002)	0.298 (.011)	0.173 (.003)

Table 6: TALD accuracy and version-relevant policy entropy H_i^{ver} on the old-version split across hint strength. Each cell reports the mean and standard error over three random seeds.

uisite: selecting the temporally applicable version of the relevant law. Existing article-prediction or citation-prediction benchmarks typically treat legal labels as static, whereas real legal reasoning must distinguish semantically related statute versions with different effective periods. Our work therefore complements downstream legal prediction tasks by diagnosing whether models can first determine which version of law should govern the case.

7.2 Temporal Legal AI

Recent work has begun to examine temporal dynamics in legal AI. ChronosLex (Santosh et al., 2024) studies temporal generalization in legal multi-label classification by training models on chronological splits. LexTempus (Santosh and Vuong, 2025) further models legal-language evolution through a dynamic mixture-of-experts framework. LexTime (Barale et al., 2025) evaluates LLMs' ability to order legal events, focusing on temporal relations within legal narratives. Closest to our motivation, LawShift (Han et al., 2025) evaluates legal judgment prediction under statutory revisions and finds that existing models struggle to adapt their judgments when underlying laws change. Our work differs in both **task** and **diagnosis**. Rather than studying temporal distribution shift, event ordering, or downstream judgment robustness under synthetic legal amendments, we directly evaluate whether LLMs can identify the temporally applicable statutory version in real judgment-derived cases. We further diagnose why models fail by separating version-centric bias, legal knowledge, general reasoning, task-specific policy alignment, and version-related reasoning entropy.

7.3 LLM Reasoning

Reasoning LLMs are achieving state-of-the-art results across a broad range of tasks that require high reasoning ability. These models are post-trained with reinforcement learning (RL) on verifiable tasks to produce an explicit chain-of-thought (CoT) before giving a final answer (Guo et al., 2025; Xu et al., 2025). However, tasks of reasoning-oriented RL training are predominantly limited to mathematics, formal logic, and competitive programming, which are structurally remote from many other domains that could be represented by legal reasoning, which requires navigating long natural-language texts, domain-specific principles, and context-sensitive expert judgment (Ma et al., 2026; Cheng et al., 2026). Whether reasoning capabilities acquired from such training can generalize to legal reasoning is, therefore, an open and practically important question. TALD offers a useful testbed for studying this question: unlike many legal reasoning tasks whose assessment is subjective or requires costly expert annotation, the version of law applicable to a given case is uniquely determined by statute and fact, making the outcome objectively verifiable and well-suited for probing model reasoning quality at scale.

8 Conclusion

We introduced TALD, a temporally grounded legal reasoning task that requires models to identify the legally applicable version of statutory law. Using a benchmark constructed from Chinese civil judgments, we showed that advanced LLMs exhibit a strong newest-law bias: they perform much better when the newest version applies, but fail severely when previous versions should govern the case. Fault-type analysis further shows that these failures are usually version-centric rather than unrelated legal citations.

Our diagnostic probes suggest that the failure is not primarily caused by missing statutory knowledge or ignorance of temporal-effect rules. Instead, stronger general reasoning does not naturally solve TALD and can even worsen old-version performance, indicating a misaligned task-specific

reasoning policy. Finally, our entropy analysis provides behavioral evidence that explicit reasoning may over-exploit a narrow newest-law trajectory, while successful correction requires reopening exploration over temporal-law alternatives. These findings highlight the need for future legal LLMs to align reasoning not only with general problem-solving ability, but also with domain-specific applicability policies.

Limitations. First, the scope of our dataset is limited to Chinese civil virdicts, where our annotators can provide reliable legal expertise; although this setting contains rich statutory revisions and well represents the Civil Law Systems, our results may not directly generalize to other legal systems. Second, we study only the temporal axis of law version. Other applicability axes, such as jurisdictional version, likewise require selecting one provision among related candidates, and whether reasoning models exhibit analogous default-policy biases along them could be a meanful next probe. Third, our mechanistic interpretation is behavioral rather than causal. Our observation of co-variation between policy entropy and TALD performance could only be based on open-weight models, and we were not able to obain observation with intervening on the training objective. Moving from behavioral evidence to a causal account requires direct intervention on the RL objective itself.

9 Ethics Statement

This work evaluates LLM behavior in legal reasoning tasks and should not be used as legal advice. Incorrect law-version determination can have serious consequences in real legal settings. We therefore emphasize that LLM outputs must be verified by qualified legal professionals. Dataset construction should follow applicable privacy and data-use rules for judgment documents, including anonymization and removal of personally identifiable information where required.

References

Claire Barale, Leslie Barrett, Vikram Sunil Bajaj, and Michael Rovatsos. 2025. LexTime: A benchmark for temporal ordering of legal events. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5220–5236.

John Bowen. 2005. *Retroactivity and the Common Law*. Hart Publishing.

Peter Chen, Xiaopeng Li, Ziniu Li, Wotao Yin, Xi Chen, and Tianyi Lin. 2026. Exploration vs exploitation: Rethinking rlvr through clipping, entropy, and spurious reward. In *International Conference on Learning Representations*.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*. Accepted to AAAI 2026.

Jorge Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Nilabjo Dey, Yonghao Zhuang, Yuheng Zha, and 1 others. 2026. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. *Advances in Neural Information Processing Systems*, 38.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

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

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Bing Han and 1 others. 2025. Lawshift: Evaluating temporal generalization of legal judgment prediction. *arXiv preprint*.

Jiuzhou Han, Paul Burgess, and Ehsan Shareghi. 2026. Legal citation prediction with llms: A comparative evaluation of instruction tuning, retrieval, and jurisdiction-specific pre-training on the auslaw citation benchmark. *Artificial Intelligence and Law*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *Preprint*, arXiv:2103.06268.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyerllama: Towards legal large language model. *arXiv preprint arXiv:2305.15062*.

Yarik Kryvoi and Shaun Matos. 2021. Non-retroactivity as a general principle of law. *Utrecht Law Review*, 17(1).

Quanzhi Li and Qiong Zhang. 2021. Court opinion generation from case fact description with legal basis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14840–14848.

Junkai Liu, Yujie Tong, Hui Huang, Bowen Zheng, Yiran Hu, Peicheng Wu, Chuan Xiao, Makoto Onizuka, Muyun Yang, and Shuyuan Zheng. 2025. Legal fact prediction: the missing piece in legal judgment prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6345–6360.

Bingfeng Luo, Yansong Feng, Jianbo Wang, Zheng Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of EMNLP*.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. 2026. General-reasoner: Advancing llm reasoning across all domains. *Advances in Neural Information Processing Systems*, 38:56596–56618.

TYSS Santosh and Tuan-Quang Vuong. 2025. Lextempus: Enhancing temporal generalizability of legal language models through dynamic mixture of experts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6608–6624.

T.Y.S.S. Santosh, Tuan-Quang Vuong, and Matthias Grabmair. 2024. Chronoslex: Time-aware incremental training for temporal generalization of legal classification tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3022–3039. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1807.02478*.

Fengli Xu, Qianyue Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, and 1 others. 2025. Toward large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*, 6(10).

Jinghan Zhang, Xiting Wang, Fengran Mo, Yeyang Zhou, Wanfu Gao, and Kunpeng Liu. 2025. Entropy-based exploration conduction for multi-step reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3895–3906, Vienna, Austria. Association for Computational Linguistics.

A System Prompts with Hint Levels

A.1 Default Prompt (No Hint)

TODO: Insert the exact base prompt used for law article prediction.

A.2 Weak Temporal-Law Hint

You are a helpful assistant. Users may ask you legal consultation questions; please respond in a factual, good-faith manner. Pay attention to the timing of events in the case and to the law's temporal applicability / intertemporal effect, especially issues of retroactivity.

A.3 Strong Temporal-Law Hint

You are a helpful assistant. Users may ask you legal consultation questions; please respond in a factual, faithful manner. Pay close attention to the timing of events in the case and to the law's temporal applicability / intertemporal effect, in particular issues of retroactivity. In civil law, the basic rule is the principle of non-retroactivity; therefore, you should carefully examine whether the relevant facts occurred at a time when the prior law should apply. That said, there are recognized exceptions, such as beneficial retroactivity, retroactive application of newly introduced provisions, and situations in which newly added specific provisions may be invoked as part of the court's judicial reasoning.