

# GLARE: Agentic Reasoning for Legal Judgment Prediction

Xinyu Yang, Chenlong Deng, Zhicheng Dou\*

Gaoling School of Artificial Intelligence, Renmin University of China

{xyysai, dou}@ruc.edu.cn

## Abstract

Legal judgment prediction serves as a pivotal task in intelligent judicial systems. Although Large Language Models (LLMs) have achieved remarkable progress in general reasoning, they struggle with tasks that require fine-grained distinctions between similar charges. These models often select plausible charges directly without discriminating among closely related alternatives. In this paper, we introduce GLARE, an agentic legal reasoning framework that enables models to actively retrieve and apply external knowledge during decision-making. Unlike static prediction, GLARE simulates comparative reasoning by dynamically expanding the decision space to include confusing candidates, then retrieving exclusionary logic from precedents and statutes to identify the correct judgment. Experiments on real-world datasets show that our method significantly outperforms strong baselines, especially on complex cases involving confusing or rare charges.

## 1 Introduction

Legal Judgment Prediction (LJP) is a core task in legal natural language processing, aiming to make accurate judgment predictions based on the factual descriptions of legal cases (Liu et al., 2023). Such predictions typically encompass three core dimensions: law articles, charges, and terms of penalty (Xu et al., 2020). This task not only provides judgment references to lawyers and judges, as well as providing legal consulting services to the general public (Luo et al., 2017; Shulayeva et al., 2017; McGinnis and Pearce, 2013).

Recently, Large Reasoning Models (LRMs) have made remarkable progress in reasoning-intensive tasks, including multi-hop question answering and strategic planning (Wang et al., 2024; Choi et al., 2025). These models can perform multi-step reasoning that mimics human thinking (Fu et al.,

2023). Intuitively, LJP appears to be an ideal fit for such models. Legal decision-making often involves comparing multiple candidate charges, evaluating whether each satisfies the legal criteria, and narrowing down to the most appropriate one based on the case facts. As a result, it is natural to expect that strong reasoning models would lead to major improvements in LJP.

However, existing reasoning models fail to deliver the expected breakthroughs in LJP. When case facts share features with multiple potential crimes, these models can effectively identify the most prominent charges and construct a reasoning chain to support them. Yet, they often neglect to further compare this choice against important similar charges or to distinguish accurately among them. This problem intensifies when alternative charges are long-tail or rare. In such cases, models tend to accept a plausible approximation while failing to extend their reasoning to the specific legal elements that would lead to a more accurate conclusion.

We argue that this failure stems from the inherent uncertainty of long-tail legal knowledge within the model. While LLMs may implicitly learn specific discriminative principles during pre-training, these fine-grained details often reside in the tail of the knowledge distribution. During the reasoning process, the model exhibits low confidence in applying these subtle rules compared to high-frequency general patterns. Consequently, the model rarely activates this dormant discriminative knowledge to reinforce its reasoning, suppressing the precise logic needed to distinguish similar charges as shown in Figure 1. Instead, it defaults to safe but imprecise conclusions that validate prominent features while overlooking critical distinctions.

To address this challenge, we propose GLARE (*AGentic Legal Reasoning FramEwork*), an agentic framework that reframes legal judgment as an active comparative reasoning process rather than passive prediction. GLARE explicitly externalizes

\*Corresponding author.

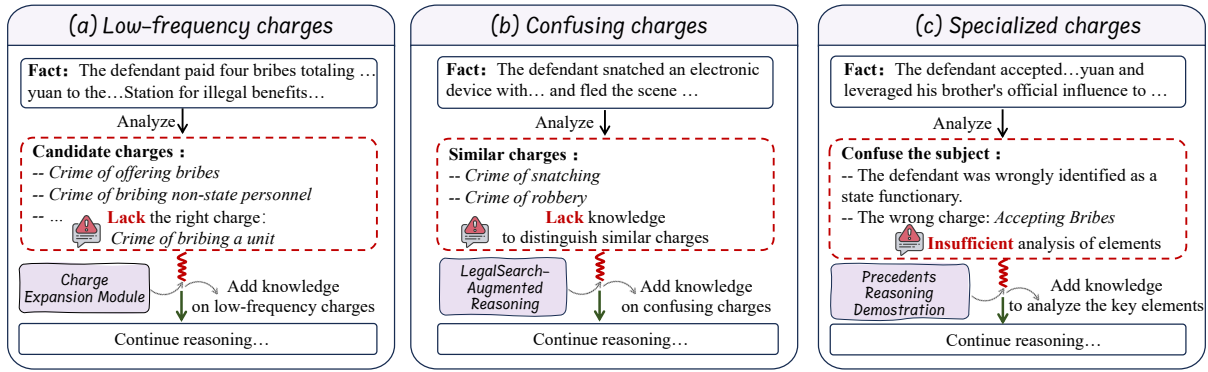


Figure 1: Lack of knowledge in three aspects: (a) Lack knowledge of low-frequency charges. (b) Lack knowledge of confusing charges. (c) Lack knowledge to analyze the key elements of the charges with strong professionalism.

the discrimination process through three integrated modules. First, the **Charge Expansion Module** (CEM) expands a diverse set of confusing charges by leveraging multiple signals, such as legal structure and historical co-occurrence. This helps the model compare a wider range of candidates and avoid premature conclusions. Second, the **Precedents Reasoning Demonstration** (PRD) module extracts discriminative reasoning patterns from similar cases, providing concrete examples that guide the model in distinguishing subtle differences. Finally, the **Legal Search-Augmented Reasoning** (LSAR) module identifies remaining ambiguities and retrieves specific legal rules when needed to resolve competing charges. We guide the model to focus its search on differences between similar charges and details of how specific laws apply, rather than general case facts. By integrating these modules, GLARE grounds the reasoning process in discriminative evidence, ensuring judgments are both logically sound and legally accurate.

Following prior work in legal judgment prediction, we conduct experiments on two publicly available real-world legal datasets. Experimental results show that our method consistently outperforms a range of strong baselines. Notably, it achieves substantial improvements on challenging cases involving confusing and difficult charges, where long-tail legal knowledge is crucial. These gains stem from our approach’s ability to effectively enrich and incorporate relevant legal knowledge.

In summary, our contributions are as follows:

(1) We introduce GLARE, an agentic framework for legal judgment prediction that enhances reasoning by dynamically integrating legal knowledge throughout the decision-making process.

(2) We design three complementary modules to

enrich the model’s reasoning process by expanding candidate charges, leveraging real-world precedents, and injecting retrieved legal knowledge.

(3) Extensive experiments on two real-world datasets show that GLARE significantly outperforms strong baselines, with especially notable gains on cases requiring crucial legal knowledge.

## 2 Related Work

**Legal judgment prediction** Legal judgment prediction has experienced significant development and has become an increasingly crucial NLP task. Earlier research (Segal, 1984) relied on artificially designed features to capture information from legal texts. Sulea et al., 2017 applied traditional machine learning methods to predict the legal judgment. Recent advances in deep learning (Xu et al., 2020; Zhang and Dou, 2023) have motivated researchers to leverage neural networks for automated text representation learning. Recently, LLMs have further promoted the progress of LJP (Deng et al., 2024a), and several studies (Wu et al., 2023; Peng and Chen, 2024) employ Retrieval-Augmented Generation (RAG) technology (Zhao et al., 2026) to enhance LLMs by incorporating external legal knowledge. However, existing LLM-based methods struggle to utilize comprehensive legal knowledge (Fei et al., 2024) and refer to the way of precedent reasoning to analyze cases. In this context, we make full use of external knowledge and precedents.

**Reasoning skills in language models** Recent work has improved LLMs’ reasoning through better prompting techniques (Sahoo et al., 2024). Wei et al. (2022) showed that chain-of-thought prompting can explicitly guide LLMs to reason step by step. In the legal domain specifically, LoT (Jiang

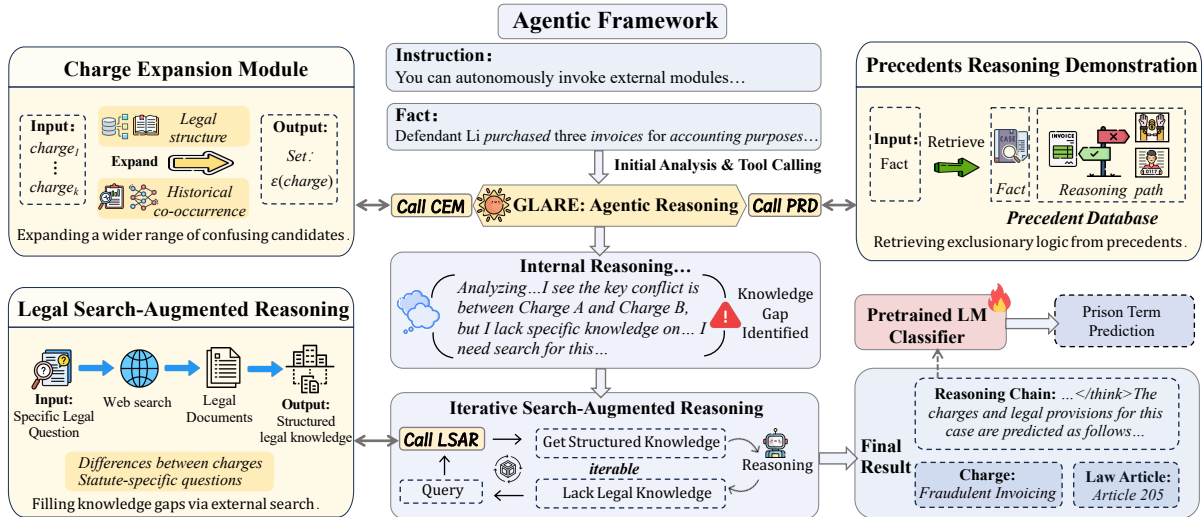


Figure 2: Overview of our agentic legal reasoning framework. LLMs autonomously expand the candidate charge space through the Charge Expansion Module, actively retrieve precedents to learn reasoning patterns, and reason over each candidate charge while iteratively invoking web search to fill knowledge gaps.

and Yang, 2023) proposed legal syllogism reasoning to improve performance on the LJP task. ADAPT (Deng et al., 2024b; Yuan et al., 2024) further established a comprehensive workflow for LJP that enables discriminative reasoning in LLMs. However, these approaches primarily rely on the LLM’s intrinsic capabilities, which inherently constrain the reasoning breadth and the depth of analysis (Zhang, 2024; Ke et al., 2025). Therefore, we propose an agentic legal reasoning framework to dynamically acquire key legal knowledge to improve the breadth and depth of reasoning.

### 3 Methodology

#### 3.1 Preliminaries

We first formally define legal judgment prediction. Given a case fact description  $f$ , the model will analyze and predict the final judgment results, including the relevant law articles, the convicted charges, and the term of imprisonment for the defendant.

In this work, we treat large language models as agentic legal reasoners that can dynamically acquire and incorporate external legal knowledge to enhance their analysis. Rather than relying solely on parametric knowledge, our approach equips the model with access to external modules, enabling it to enrich its reasoning with case-specific legal context. Given a case fact  $f$  and a set of external modules  $M$ , the model performs step-by-step analysis to construct a coherent reasoning chain  $R$  and arrive at a final judgment prediction  $p$ . We formalize this process as a mapping:  $(f, \mathcal{M}) \rightarrow (R, p)$ .

#### 3.2 Agentic Legal Reasoning Framework

We propose GLARE, an agentic legal reasoning framework that autonomously invokes external modules to support comprehensive and informed judgment prediction. As shown in Figure 2, GLARE follows a structured three-stage reasoning pipeline:

- 1. Charge Expansion:** The model begins by analyzing the fact and generating preliminary candidate charges. To prevent premature narrowing of the decision space, it autonomously triggers the Charge Expansion Module to supplement the initial candidates with legally similar charges.
- 2. Precedent-Enhanced Reasoning:** The model retrieves relevant precedents from an offline-constructed database that includes fact descriptions and synthesized reasoning chains. The reasoning chains were constructed in advance to illustrate the key distinctions between confusing charges. These precedents serve as case-specific reasoning demonstrations, helping the model better understand how similar legal criteria apply and guiding it through more precise reasoning via in-context learning.
- 3. Iterative Search-augmented Reasoning:** As the model reasons through each candidate charge, it dynamically identifies knowledge gaps such as missing legal definitions and charge-specific thresholds. Rather than treating retrieval as a one-time step, the model interleaves reasoning and

retrieval in a loop. Retrieved results are injected back into the reasoning context, enabling the model to refine its current analysis. This iterative process continues until the model has collected sufficient knowledge to complete its reasoning and reach a final judgment.

The three modules collaboratively supplement legal knowledge and extend the legal reasoning chain. Next, we will introduce these three modules.

### 3.3 Charge Expansion Module

To enable charge comparison and avoid premature conclusions, we expand each candidate charge by retrieving related charges. The expansion is based on two complementary perspectives: legal structure and historical co-occurrence.

**Legal Structure-based Expansion.** The Criminal Law is organized into chapters, each representing a specific legal interest or domain. Charges within the same chapter typically differ in subtle legal criteria, while charges across different chapters may involve similar actions or consequences but fall under distinct legal categories. To capture both fine-grained intra-domain distinctions and cross-domain conceptual similarities, we retrieve related charges from both within the same chapter and across different chapters.

Specifically, for a given charge  $c$ , we use the pretrained dense retriever BGE (Xiao et al., 2024) to find the top- $k$  most similar charges from (a) the same chapter and (b) other chapters:

$$\mathcal{E}_1(c) = \text{topk}_{\text{same}}(c) \cup \text{topk}_{\text{diff}}(c), \quad (1)$$

where  $\text{top-}k_{\text{same}}(c)$  and  $\text{top-}k_{\text{diff}}(c)$  represent the most similar charges from the same and different chapters, respectively. This dual-source expansion helps the model compare similar alternatives, reducing the risk of overlooking relevant charges.

**History-based Expansion.** Certain charges tend to appear together in real-world cases, reflecting practical legal dependencies or common joint indictments. We leverage the MultiLJP (Lyu et al., 2023) dataset, where each case may involve multiple defendants and multiple charges. By analyzing these cases, we construct a co-occurrence dictionary that records how frequently each pair of charges appears together. For a given charge  $c$ , we select the top- $k$  most frequently co-occurring charges as the expansion set  $\mathcal{E}_2(c)$ .

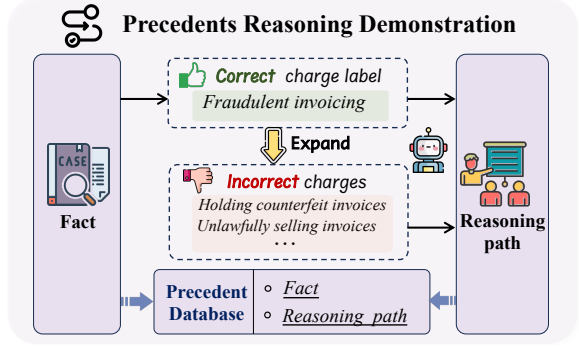


Figure 3: The module of Precedents Reasoning Demonstration: LLM analyzes the reasons for the selection or exclusion of each charge based on facts, thereby generating the reasoning path of precedents.

**Final Expansion Set.** Given an initial set of candidate charges  $\{c_1, c_2, \dots\}$  predicted by the language model, we apply the two strategies above to expand each charge:

$$\mathcal{E}(c_i) = \mathcal{E}_1(c_i) \cup \mathcal{E}_2(c_i). \quad (2)$$

### 3.4 Precedents Reasoning Demonstration

Previous precedent-based approaches (Wu et al., 2023; Chen and Zhang, 2025; Santosh et al., 2024) typically retrieve the fact description and final judgment of prior cases, then insert them directly into the prompt. However, such methods offer little insight into the *reasoning process* behind those decisions. As a result, they tend to rely on shallow fact matching rather than learning how to distinguish between legally similar charges.

Therefore, we construct reasoning-augmented precedents that make the decision logic explicit. The precedent database is built from the CAIL2018 and CMDL training sets, comprising 163,563 carefully curated cases spanning 190 charges and covering both single-defendant and multi-defendant scenarios. We perform systematic filtering, deduplication, and quality control to ensure fairness, and the database is continuously updated to adapt to the evolving judicial environment.

As shown in Figure 3, we first expand its original charge  $c$  into a set of similar charges  $C$ . Given the case fact  $f$ , the correct charge  $c$ , and the set of alternatives  $C$ , we prompt LLM to generate a reasoning path  $r$  that explains why  $c$  is appropriate and why the other candidates in  $C \setminus \{c\}$  should be excluded.<sup>1</sup> This reasoning is generated **offline** and stored together with the case facts.

<sup>1</sup>We provide the detailed descriptions in Appendix C.

### 3.5 Legal Search-Augmented Reasoning

While recent retrieval-augmented generation (RAG) approaches (Wu et al., 2023; Peng and Chen, 2024; Feng et al., 2024) enhance legal models by retrieving precedents, statutes, and charge definitions, they remain limited in key aspects. Specifically, they often fail to resolve fine-grained distinctions between similar charges or provide detailed rules to determine facts. Moreover, these methods rely on static retrieval from fixed knowledge bases, making them inflexible and unable to accommodate evolving judicial practices.

To address these limitations, we introduce a **dynamic and iterative legal search-augmented reasoning mechanism**. Rather than passively injecting generic legal content, our method allows the LLM to actively identify knowledge gaps during reasoning and generate targeted queries. These queries focus on *subtle differences between candidate charges* and *statute-specific questions*. The system retrieves relevant legal texts from the web in real time, enabling up-to-date and context-related augmentation. We also introduce dedicated mechanisms to control noise and mitigate potential risks of data leakage, as detailed in Appendix D.

We further ground the model’s reasoning in a **sylogistic structure**: the retrieved legal context serves as the major premise, the fact as the minor premise, and the conclusion is derived through logical alignment (Jiang and Yang, 2023). This helps the model remain grounded in factual evidence and reduce hallucinations. The overall reasoning process is formalized as an iterative function:

$$R_t = f_{\theta}(R_{<t}, q_t, d_t, f), \quad (3)$$

where  $R_t$  denotes the current reasoning state,  $R_{<t}$  are the historical reasoning paths,  $q_t$  and  $d_t$  are the query and corresponding retrieved documents of this step, and  $f$  is the case fact.

This design enables the model to incrementally construct a legally grounded reasoning chain, adaptively integrating external knowledge as needed. By decoupling retrieval from static knowledge bases and aligning it with the model’s evolving needs, our framework offers greater flexibility to real-world legal dynamics.

### 3.6 Sentence Term Prediction

As sentencing prediction is highly discretionary and current LLMs still struggle to generate reliable quantitative terms, we rely on a conventional fine-tuned classifier for this stage. Concretely, once

Dataset	CAIL2018	CMDL
# Train cases	100,531	63,032
# Test cases	1,000	834
# Charges	190	164
# Articles	175	147
# Average criminal per case	1	3.79
Average length per case	409.6	1124.94

Table 1: Statistics of the datasets used in the paper.

the charge and law article have been determined, we concatenate the reasoning chains with the case facts and fine-tune a pre-trained language model to predict the penalty term. See Appendix E for implementation details.

## 4 Experiments

### 4.1 Datasets and Evaluation

We conduct experiments in both single-defendant and multi-defendant scenarios to verify the effectiveness of our method in practical applications. For the single-defendant case, we use the CAIL2018 dataset (Xiao et al., 2018). For the multi-defendant case, we adopt the CMDL dataset (Huang et al., 2024). We uniformly sample across all charges to construct a balanced test set. The details are shown in Table 1. For evaluation metrics, we adopt the same measures used in prior work: Accuracy (Acc.), Macro Precision (Ma-P), Macro Recall (Ma-R), and Macro F1 (Ma-F).

### 4.2 Baselines

We compare our method against two categories of baseline approaches:

**Classification Methods:** These methods take legal judgment prediction as a classification task, relying on supervised learning with labeled datasets. **TopJudge** (Zhong et al., 2018) employs a graph structure to model the topological dependency among the three subtasks: charge prediction, law article prediction, and sentence term prediction. **NeurJudge** (Yue et al., 2021) integrates a legal knowledge graph into the neural architecture, capturing explicit relationships among legal entities and improving reasoning over structured legal knowledge. **BERT** (Devlin et al., 2019), a standard pre-trained transformer model, is adapted to legal texts via supervised training. It serves as a strong baseline for judgment prediction tasks. **Lawformer** (Xiao et al., 2021) is built upon Longformer (Beltagy et al., 2020) and further pretrained

Methods	Charge				Law Article				Prison Term			
	Acc.	Ma-P	Ma-R	Ma-F	Acc.	Ma-P	Ma-R	Ma-F	Acc.	Ma-P	Ma-R	Ma-F
<i>Classification Methods</i>												
TopJudge	52.1	50.9	45.7	43.5	52.8	47.7	43.8	41.2	33.6	28.4	25.5	25.0
LADAN	76.7	73.4	71.0	69.5	77.5	71.0	69.2	67.5	35.8	30.6	25.7	25.2
NeurJudge	77.1	79.3	73.5	73.4	78.7	78.0	74.1	73.8	36.3	35.6	36.1	35.5
BERT	84.0	79.5	81.1	78.6	84.6	77.7	80.2	77.0	37.7	29.7	32.0	28.1
Lawformer	71.3	58.2	62.7	57.8	72.9	58.1	61.4	56.9	36.5	30.0	28.5	27.4
<i>Direct Reasoning</i>												
Qwen2.5-32B	74.5	75.3	69.3	69.1	77.1	73.3	66.6	67.1	24.5	27.1	16.7	15.8
QwQ-32B	82.5	86.9	80.5	80.9	84.0	83.1	76.1	77.0	20.3	30.5	23.4	18.4
Qwen2.5-72B	76.6	78.9	72.2	72.3	77.7	73.4	66.8	67.3	20.3	33.4	22.2	17.6
DeepSeek-V3	79.5	83.4	76.3	77.0	80.0	79.3	70.7	72.0	30.6	25.7	20.6	19.7
DeepSeek-R1-671B	84.8	86.3	81.3	81.7	87.2	86.8	81.8	82.6	37.8	33.8	30.1	31.0
<i>Retrieval-augmented Reasoning</i>												
Precedent-based-RAG												
Qwen2.5-32B	88.5	88.2	85.8	85.7	89.4	87.2	83.7	84.5	27.4	32.1	20.8	18.4
QwQ-32B	89.4	89.9	87.3	87.1	90.4	88.4	85.2	85.4	28.3	21.8	22.2	21.5
Qwen2.5-72B	88.1	87.5	85.1	84.9	89.4	86.8	83.9	84.0	34.3	33.9	27.4	26.1
PLJP-QwQ-32B	84.3	85.6	81.6	81.6	84.4	84.1	79.1	79.6	36.1	36.9	34.2	33.2
Search-o1-QwQ-32B	81.8	85.3	78.8	79.3	83.9	83.3	76.4	77.4	19.8	34.2	24.4	18.6
<i>Agentic Retrieval-augmented Reasoning</i>												
GLARE-Qwen2.5-32B	<b>89.8</b>	89.8	87.8	87.8	90.4	89.2	87.3	87.5	39.3	35.0	34.9	34.3
GLARE-QwQ-32B	89.7	<b>90.7</b>	<b>88.6</b>	<b>88.6</b>	<b>91.3</b>	<b>90.6</b>	<b>88.3</b>	<b>88.5</b>	<b>40.3</b>	<b>37.8</b>	<b>39.8</b>	<b>38.2</b>

Table 2: Performance comparison on the CAIL2018 dataset. The best results are in bold.

on large-scale Chinese legal corpora, which enhances its ability to process longer legal documents and capture complex contextual semantics.

**LLM-based Methods:** These methods utilize LLMs to perform legal reasoning in zero-shot or few-shot settings (Brown et al., 2020). **Direct Reasoning** directly feeds the case facts into the LLM to predict the final judgments, without relying on any retrieval augmentation or additional external context. The models used in this setting include Qwen2.5-32B/72B-Instruct (Yang et al., 2024), QwQ-32B (Team, 2025), and DeepSeek-R1-671B (DeepSeek-AI, 2025). **Retrieval-augmented Reasoning:** (1) *Precedent-based RAG* enhances reasoning by retrieving top-5 precedents, including their facts and labels, which are appended to the prompt. The models used in this setting include Qwen2.5-32B/72B-Instruct, QwQ-32B. (2) *Search-o1* (Li et al., 2025) dynamically retrieves external knowledge when it encounters uncertain or ambiguous knowledge in the general domain. We use the reasoning model QwQ-32B in this setting. (3) *PLJP* (Wu et al., 2023) leverages the strength of both LLM and domain models in the context of precedents. We take the BERT as the predictive model, and take the QwQ-32B as the implementation of the LLM. Implementation details are shown in Appendix E.

### 4.3 Experiment Settings

In our experiments, we adopt Qwen2.5-32B (Yang et al., 2024) and QwQ-32B (Team, 2025) as the base models to run the full reasoning pipeline. For generation, we set the following parameters: a maximum of 32,768 tokens and a temperature of 0.6. For charge expansion, we set the top- $k$  expanded charges to 3 in each expansion method. For precedent retrieval, we use SAILER (Li et al., 2023) to encode case facts and set the top- $k$  retrieved precedents to 5. In the legal search module, we utilize Serper API<sup>2</sup> with the region configured for China and the number of returned results limited to the top 10. For charges that are not in the predefined label set, we map them to the most similar charge within the label set using BGE (Xiao et al., 2024).

### 4.4 Experiment Results

The results are reported in Table 2 and Appendix F, and next we will analyze the experimental results:

**1. Our method has demonstrated consistent performance improvements in all three tasks,** highlighting the effectiveness of our comparative reasoning process. Unlike passive prediction approaches such as Direct Reasoning and Retrieval-augmented Reasoning, our agentic method explicitly externalizes the discrimination process through

<sup>2</sup><https://serper.dev>

Methods	CAIL2018				CMDL			
	Charge		Law Article		Charge		Law Article	
	Acc.	Ma-F	Acc.	Ma-F	Acc.	Ma-F	Acc.	Ma-F
<i>Direct Reasoning</i>								
Qwen2.5-32B	60.2	39.3	63.7	41.2	57.4	64.7	57.9	63.7
QwQ-32B	78.4	57.0	79.2	58.2	67.9	72.9	69.8	74.2
<i>Retrieval-augmented Reasoning</i>								
Precedent-based-RAG-Qwen2.5-32B	82.6	62.7	83.0	62.3	65.7	69.5	65.3	67.6
Precedent-based-RAG-QwQ-32B	84.6	65.5	84.6	67.6	72.8	74.8	71.3	73.2
<i>Agentic Retrieval-augmented Reasoning</i>								
GLARE-Qwen2.5-32B(ours)	86.9	68.6	86.5	68.3	73.5	75.5	71.9	73.4
GLARE-QwQ-32B(ours)	<b>90.7</b>	<b>75.7</b>	<b>91.1</b>	<b>75.4</b>	<b>76.0</b>	<b>79.5</b>	<b>74.0</b>	<b>76.7</b>

Table 3: Performance comparison on difficult charges.

three integrated modules, enabling fine-grained discrimination between similar charges. Compared to the Direct Reasoning setting, our method improves charge prediction by 7.7% and law article prediction by 11.5% in F1 score. Appendix H discusses how our method enhances LLM reasoning beyond knowledge supplementation.

**2. Large reasoning models such as QwQ-32B and DeepSeek-R1 outperform instruct models,** indicating that LJP inherently requires multi-step reasoning. While Retrieval-augmented Reasoning leverages precedents for similarity matching and increases the likelihood that the correct charge appears in the decision space, it still lacks genuine discriminative reasoning. PLJP expands the decision space by predicting top-k candidate labels with a domain model, but it fails to provide sufficient discriminative knowledge to support reasoning. Search-o1 introduces dynamic external retrieval, but its reasoning process still lacks explicit comparative logic. In contrast, our method simulates comparative reasoning and explicitly grounds the reasoning process in discriminative evidence.

**3. Classification methods outperform LLM-based methods in prison term prediction.** As sentencing prediction is highly discretionary, LLMs have difficulty producing stable quantitative terms. Our method preserves the legal reasoning capability of LLMs while leveraging the robustness of conventional pre-trained classifiers to generate trustworthy prison term predictions.

#### 4.5 Ablation Study

To evaluate the effectiveness of each component in GLARE, we conducted the following ablation experiments: **(1) w/o CEM:** The Charge Expansion Module is removed, so the model cannot expand

Methods	Charge		Law Article	
	Acc.	Ma-F	Acc.	Ma-F
w/o CEM	89.6	87.7	90.3	85.2
w/o PRD	80.0	78.1	81.6	75.4
w/o LSAR	89.6	87.9	90.4	86.5
GLARE(ours)	<b>89.7</b>	<b>88.6</b>	<b>91.3</b>	<b>88.5</b>

Table 4: Ablation Study. The best results are in bold.

a diverse set of candidate charges. **(2) w/o PRD:** The Precedents Reasoning Demonstration module is removed, so the model cannot refer to the reasoning path from precedents. **(3) w/o LSAR:** The Legal Search-Augmented Reasoning module is removed, disabling the model’s ability to supplement its knowledge via external legal search when faced with ambiguous or unfamiliar charges.

As shown in Table 4, the removal of any single module results in degraded performance. In particular, removing PRD causes the most significant degradation: the accuracy of charge prediction drops from 89.7% to 80%. This highlights the importance of precedent-guided comparative reasoning in LJP tasks. Removing CEM weakens the model’s ability to expand a diverse set of confusing charges, while LSAR helps to fill knowledge gaps by retrieving authoritative legal information.

#### 4.6 Efficiency Analysis

Legal analysis needs multi-step reasoning and slow thinking, so the latency is less important. Nevertheless, we still analyze to further understand each module. As shown in Figure 5 (a), we can find that: *(1) the overall inference efficiency is relatively high.* The system is well-balanced, with an average of 5.17 reasoning rounds per case and 1.7-1.8 calls per module, showing no major redundancy. *(2) The*

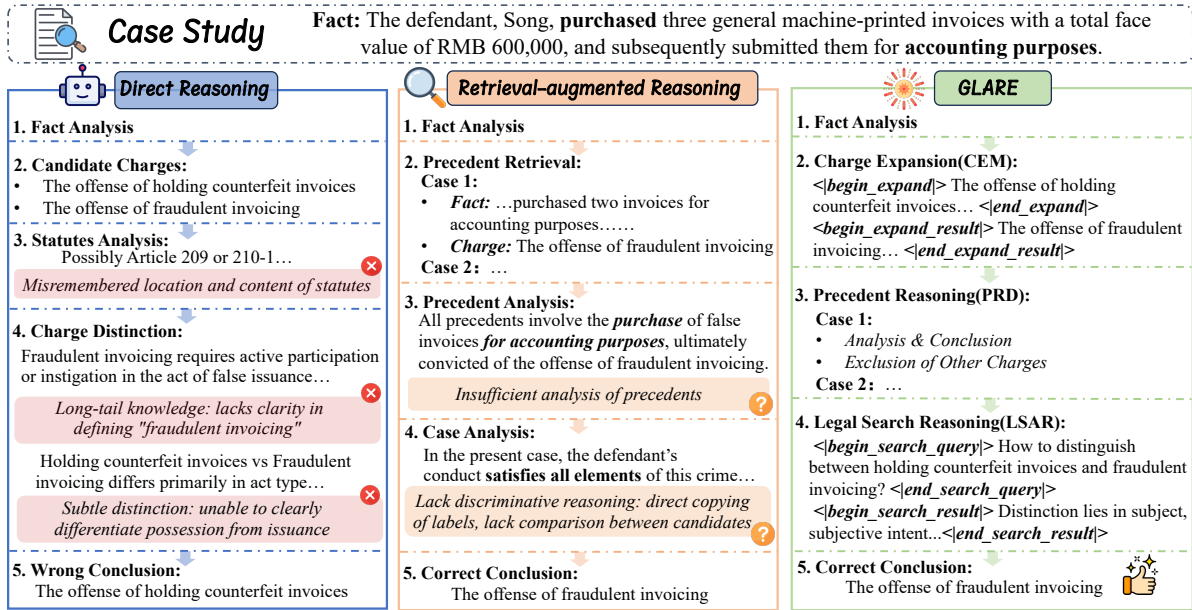


Figure 4: Case Study. The red part highlights the model’s limitations due to insufficient long-tail knowledge, while the yellow part demonstrates the lack of discriminative reasoning in vanilla precedent-based RAG reasoning.

CEM module is the most efficient. This allows for a more comprehensive charge analysis compared to direct reasoning or RAG as shown in the figure 5 (b). (3) The PRD module has the highest latency, but within an acceptable range.

#### 4.7 Case Study

As shown in Figure 4, we conducted a case study on three LLM-based methods. Direct Reasoning relies on the LLM’s internal knowledge. However, fine-grained discriminative details often reside in the tail of the knowledge distribution, and LLMs rarely activate such knowledge to reinforce their reasoning. RAG-based methods often offer little insight into the reasoning process behind those decisions, so the model does not actually learn how to distinguish between similar charges. Our method enriches the model’s reasoning process through three complementary modules, enabling more explicit and reliable comparative analysis. We also conducted a human evaluation in Appendix G.

#### 4.8 Performance on Difficult Charges

To evaluate GLARE’s ability to handle challenging charges requiring long-tail knowledge, we conducted experiments on low-frequency charges with less than 100 cases (e.g., the crime of bribing a unit) and confusing charges (e.g., robbery vs. snatching). The results are reported in Table 3, which shows that our method outperforms Direct Reasoning by over 10% and Retrieval-augmented Reasoning by

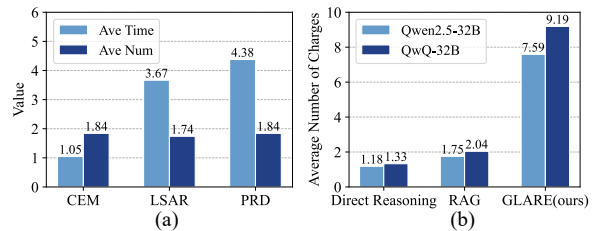


Figure 5: (a) Efficiency analysis of each module. (b) Average charge numbers of different methods.

over 5%. RAG-based methods struggle to retrieve relevant precedents for such charges, leading to poor performance, while direct reasoning lacks long-tail knowledge. These results highlight the strength of our external modules in supplementing legal reasoning with critical knowledge.

## 5 Conclusion

In this study, we propose an agentic framework, GLARE, to address the limited discriminative reasoning of LLMs in legal judgment prediction. GLARE employs three complementary modules to dynamically integrate legal knowledge throughout the decision-making process and enrich the model’s reasoning capability. Experiments on two real-world datasets show that our method significantly outperforms strong baselines. We believe that GLARE holds great potential for real-world legal applications and will contribute meaningfully to the advancement of intelligent judicial systems.

## Limitations

**Generalizability** We adopted the legal dataset from China Judgments Online to verify the applicability of the method in China’s judicial system. However, the GLARE framework is applicable to countries following both common law and civil law systems. When applied to the actual judicial practice of a specific country, we need to inject the specific legal knowledge base of each country and adapt to the local judicial culture.

**Efficiency** Our method promotes the reasoning ability of the model through multiple rounds of reasoning and the invocation of three modules. Although this process has an increased time cost compared to the traditional direct reasoning method, the task of legal judgment prediction itself is a task that requires multi-step reasoning and slow thinking. Moreover, this time cost is much less than the time needed for humans to analyze cases in real life. Therefore, such a time cost is acceptable.

**Noise Control** Although the charge expansion module may introduce noise in the process of generating rich candidate charges, our experiments show that the influence of the noise on the final judgment is effectively controlled by the PRD and LSAR modules. The PRD module retrieves the most relevant precedents and provides detailed distinctions between similar charges. The LSAR module ensures the retrieval quality through a dual constraint mechanism: content constraints prevent data leakage caused by directly retrieving the case fact, while source constraints guarantee authority by retrieving exclusively from official channels.

**Prison Term Discretization** In real-world judicial practice, sentencing is influenced by numerous factors, including case circumstances, defendant background, and mitigating or aggravating conditions. Many situations cannot be exhaustively specified and ultimately rely on judicial discretion. Given this inherent complexity, we follow prior work and evaluate sentencing prediction by mapping continuous predictions to discrete intervals. However, we acknowledge that this approach may obscure quantitative errors and that continuous prediction better aligns with real judicial needs. As models and datasets improve, moving beyond interval-based evaluation remains an important future direction.

## Ethical Discussion

**Potential Bias in Legal Data** Large language models may learn historical bias from legal judgments in training data. In practice, judicial decisions may be influenced by many external factors, such as public opinion, regional differences or the personal inclinations of judges. We need to identify possible biases before deploying such models in real-world scenarios.

**Human-Centric Deployment** Our system is designed to assist judges by providing supplementary recommendations rather than replacing human decision-making. We advise users to critically evaluate the model’s predictions and make independent decisions about their adoption, rather than uncritically accepting the model’s reasoning.

## Acknowledgements

This work was supported by National Natural Science Foundation of China No. 62272467. The work was partially done at the Beijing Key Laboratory of Research on Large Models and Intelligent Governance.

## References

- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#). *CoRR*, abs/2004.05150.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Quan Ze Chen and Amy Xian Zhang. 2025. [Case law grounding: Using precedents to align decision-making for humans and AI](#). In *Proceedings of the ACM Collective Intelligence Conference, CI 2025, San Diego, CA, USA, August 4-6, 2025*, pages 226–238. ACM.
- Daewon Choi, Jimin Lee, Jihoon Tack, Woomin Song, Saket Dingliwal, Sai Muralidhar Jayanthi, Bhavana Ganesh, Jinwoo Shin, Aram Galstyan, and Sravan Babu Bodapati. 2025. [Think clearly: Improving reasoning via redundant token pruning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 21437–21451. Association for Computational Linguistics.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.

- Chenlong Deng, Kelong Mao, and Zhicheng Dou. 2024a. [Learning interpretable legal case retrieval via knowledge-guided case reformulation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1253–1265. Association for Computational Linguistics.
- Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. 2024b. [Enabling discriminative reasoning in llms for legal judgment prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 784–796. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. [Lawbench: Benchmarking legal knowledge of large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7933–7962. Association for Computational Linguistics.
- Jianzhou Feng, Lazhi Zhao, Haonan Qin, Yiming Xu, and Ziqi Wang. 2024. Cadlra: A multi-charge prediction method based on the criminal act-driven law retrieval augmentation. *Engineering Applications of Artificial Intelligence*, 134:108619.
- Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. [Complexity-based prompting for multi-step reasoning](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- 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.
- Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the nineteenth international conference on artificial intelligence and law*, pages 417–421.
- Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, Caiming Xiong, and Shafiq Joty. 2025. [A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems](#). *Trans. Mach. Learn. Res.*, 2025.
- Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. Sailer: structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1035–1044.
- Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. [Search-o1: Agentic search-enhanced large reasoning models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 5420–5438. Association for Computational Linguistics.
- Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. [ML-ljp: Multi-law aware legal judgment prediction](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 1023–1034, New York, NY, USA. Association for Computing Machinery.
- Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. [Learning to predict charges for criminal cases with legal basis](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2727–2736. Association for Computational Linguistics.
- Yougang Lyu, Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang, and Zhaochun Ren. 2023. [Multi-defendant legal judgment prediction via hierarchical reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 2198–2209. Association for Computational Linguistics.
- John O McGinnis and Russell G Pearce. 2013. The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services. *Fordham L. Rev.*, 82:3041.
- Xiao Peng and Liang Chen. 2024. [Athena: Retrieval-augmented legal judgment prediction with large language models](#). *CoRR*, abs/2410.11195.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. [A systematic survey of prompt engineering in large language models: Techniques and applications](#). *CoRR*, abs/2402.07927.
- T. Y. S. S. Santosh, Mohamed Hesham Elganayni, Stanislaw Sójka, and Matthias Grabmair. 2024. [Incorporating precedents for legal judgement prediction on european court of human rights cases](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November*

- 12-16, 2024, Findings of ACL, pages 3743–3750. Association for Computational Linguistics.
- Jeffrey A Segal. 1984. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900.
- Olga Shulayeva, Advait Siddharthan, and Adam Wyner. 2017. Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, 25(1):107–126.
- Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith. 2017. Exploring the use of text classification in the legal domain. In *Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017), London, UK, June 16, 2017*, CEUR Workshop Proceedings. CEUR-WS.org.
- Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.
- Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *CoRR*, abs/2401.06805.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12060–12075. Association for Computational Linguistics.
- Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.
- Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.
- Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3086–3095. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Tianqianjin Lin, Pengwei Yan, Changlong Sun, and Xiaozhong Liu. 2024. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 7577–7597. Association for Computational Linguistics.
- Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 973–982.
- Han Zhang and Zhicheng Dou. 2023. Case retrieval for legal judgment prediction in legal artificial intelligence. In *China National Conference on Chinese Computational Linguistics*, pages 434–448. Springer.
- Jianqiu Zhang. 2024. Should we fear large language models? A structural analysis of the human reasoning system for elucidating LLM capabilities and risks through the lens of heidegger’s philosophy. *CoRR*, abs/2403.03288.
- Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2026. Retrieval-augmented generation for ai-generated content: A survey. *Data Sci. Eng.*, 11(1):1–29.
- Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3540–3549.

## Appendix

### A Instructions for Invoking External Modules

(a) Charge Expansion Module and Precedents Reasoning Demonstration: We instruct LLMs to generate preliminary candidate charges enclosed within special symbols `<begin_expand>` and `<end_expand>`. When such symbols are detected, LLMs will stop reasoning and invoke the Charge Expansion Module to expand charges. The expanded charges are injected back into the reasoning chain along with retrieved precedents. (b) Legal Search-Augmented Reasoning: Similarly, we instruct LLMs to generate queries encapsulated between special symbols `<begin_search>` and `<end_search>` to trigger the retrieval mechanism. The system then invokes a web search to obtain relevant information. Detailed instructions are shown in Figure 7.

### B Prompts of GLARE

The main analysis process is: (1) LLMs initially analyze the case and generate preliminary candidate charges, then expand a diverse set of similar charges; (2) retrieve the most relevant precedents; (3) integrate the retrieved precedents and the expanded charges into the reasoning chain together, and then analyze each candidate charge via in-context learning; (4) during the reasoning process, when encountering knowledge gaps, autonomously acquire structured legal knowledge through web search, ultimately outputting applicable law articles and convicted charges. Detailed prompts are shown in Figure 8.

### C Detailed Descriptions of Precedents Reasoning Demonstration

We adopt the training sets of the CAIL2018 and CMDL datasets as the initial precedent database, and apply systematic filtering, deduplication, and quality control. In particular, we remove cases that are highly similar to the test set or contain incomplete information, and ensure balanced coverage across diverse case types. This design enhances the diversity and generalization ability of the precedent database.

For the generation of reasoning paths, we first expand the original charge label into a set of similar charges by Charge Expansion Module. Then given the case fact  $f$ , the correct charge  $c$ , and the

#### Precedents Reasoning Demonstration

Given a description of a legal case along with the correct charge, as well as other potential charges, analyze the reasons for selecting the correct charge while excluding the others.

1. Case description: {fact}
2. Correct charge: {charge\_label}
3. Other charges: {expanded\_crimes}

Figure 6: The prompt used for PRD.

set of alternatives  $C$ , we prompt Qwen2.5-72B-Instruct (Yang et al., 2024) to generate a reasoning path  $r$ . Detailed prompts are shown in Figure 6.

To ensure that the synthesized reasoning paths are legally sound and aligned with real-world judicial practice, we employ a multi-stage quality control pipeline during data construction. First, we automatically filter out low-quality cases using rule-based detectors, including checks for hallucinated facts, unsupported legal conclusions, contradictions with the final judgment, and missing comparisons with alternative charges. Second, a subset of precedents is randomly sampled for human verification by postgraduate law students under expert supervision, and the resulting feedback is iteratively incorporated to refine the prompting template and filtering rules. Finally, we evaluate the internal consistency of the generated reasoning by prompting another LLM to act as a critic and score whether each reasoning path is logically coherent, fact-grounded, and norm-consistent. Cases that fail any validation step are discarded. Through this pipeline, we maintain a high-quality precedent base in which the reasoning not only states the correct conclusion but also reflects the substantive legal logic used to distinguish between similar charges.

Next, we present a specific example of a precedent reasoning path:

**Case Description:** *From Nov 2011 to May 2012, defendant Song purchased 2 forged tax invoices (total ¥441,563) to document rent payments from Qingdao Gaoshi Thread Co., Ltd. Forensic examination confirmed all invoices were counterfeit.*

**Analysis and Conclusion:** *The correct Charge: Issuing False Invoices (Article 205-1 of Criminal Law). The defendant’s conduct satisfies the essential elements of issuing false invoices: According to Article 205-1, this crime involves illegally issuing*

ordinary invoices for others, for oneself, having others issue invoices for oneself, or introducing others to issue invoices. Song purchased forged invoices and provided them to the company for accounting purposes, constituting typical "issuing false invoices" behavior. The intent is not merely to possess counterfeit invoices but to fabricate transaction records for financial fraud.

**Exclusion of Other Charges:** (1) **Possession of Forged Invoices:** While Song did possess forged invoices, this was merely the means to commit invoice fraud. According to the Doctrine of Implied Offense in criminal theory, when one act violates multiple charges, the more severe offense applies. Issuing false invoices carries heavier penalties (maximum 7-year imprisonment) than possession (maximum 2 years). (2) **Issuing Special VAT Invoices:** Only applies to VAT invoices or other tax-deductible invoices (Article 205). The case involves ordinary invoices without tax-related functions. (3) **Financial Instrument Fraud:** Requires intent to illegally possess others' property through forged instruments (Article 194). Song aimed to document rent payments, not defraud directly. (4) **False Registered Capital Reporting:** Concerns fraudulent capital contributions during company establishment (Article 158). Irrelevant to invoice-related conduct. (5) **Illegal Sale of Invoices:** Applies to selling genuine invoices (Article 207). Song purchased rather than sold invoices.

**Conclusion:** Defendant Song's conduct constitutes **Issuing False Invoices** (Article 205-1) as it fulfilled all statutory elements with greater social harm than alternative charges. Other charges were excluded due to: (1) mismatched conduct objects, (2) different subjective intents, or (3) being secondary implicated offenses.

## D Dual-Constraint Mechanism for LSAR

To further prevent potential data leakage risks that may arise from web-based retrieval, we design a dual-constraint mechanism consisting of **content constraints** and **source constraints**.

First, we avoid using any direct descriptions of case facts as search queries. Instead, the LLM generates generalized legal questions that focus on subtle distinctions between candidate charges and abstract legal principles applicable to the scenario. All retrieved documents are further screened to exclude any text containing specific details of ongoing or historical cases, thereby eliminating the

risk of exposing raw case data.

Second, source constraints ensure the authoritativeness and compliance of retrieved materials. We exclusively source legal texts and interpretations from official and publicly-accessible judicial or governmental platforms, such as the China Courts Network, the Supreme People's Court portal, and other certified statutory databases. No retrieval from commercial legal platforms, private datasets, or case-level repositories is allowed.

Through this dual-constraint design, the LSAR module benefits from real-time legal knowledge while strictly eliminating any possibility that the retrieval process leaks case-specific data or injects external case facts into the model's reasoning.

## E Implementaion Details

We adopt a BERT-based classifier for sentence term prediction. The model is trained on the training splits of the CAIL2018 and CMDL datasets. For each case, we first run our GLARE pipeline to obtain the reasoning chains of law articles and charges. The model input is constructed by concatenating the case facts with the reasoning chains in the format: `[CLS] Fact [SEP] Reasoning Chain [SEP]`. Following prior work, we discretize the sentence term into 11 intervals and train the classifier to predict the interval label. The learning rate is set to  $2.0e-5$ , and the number of training epochs is 20.

**Search-o1** is a general framework that enhances large reasoning models (LRMs) with an agentic retrieval-augmented generation mechanism. When applied in the legal domain, we prompt QwQ-32B to autonomously generate queries related to the legal issues under discussion. Meanwhile, we impose both content-level and source-level constraints to prevent potential data leakage and ensure the reliability of the retrieved information. In practice, we use the Serper API with the region configured for China and restrict the number of returned results to the top 10.

The core objective of Search-o1 is knowledge supplementation. In contrast, our method places greater emphasis on comparative reasoning. Specifically, we design a charge-expansion mechanism to include confusing candidate charges and innovatively introduce the reasoning trajectories of precedents, enabling the model to learn discriminative reasoning patterns. As a result, our approach achieves superior performance in legal judgment prediction tasks.

Methods	Charge				Law Article				Prison Term			
	Acc.	Ma-P	Ma-R	Ma-F	Acc.	Ma-P	Ma-R	Ma-F	Acc.	Ma-P	Ma-R	Ma-F
<b>Direct Reasoning</b>												
Qwen2.5-32B	70.4	76.0	80.2	77.2	73.5	79.3	81.0	79.3	19.4	22.5	18.5	15.7
QwQ-32B	77.9	81.7	83.7	82.1	79.5	83.4	85.1	83.5	22.3	26.6	24.4	19.5
Qwen2.5-72B	73.6	77.5	80.6	78.5	76.1	80.1	80.9	79.8	22.9	28.4	24.0	19.5
DeepSeek-R1-671B	81.4	84.1	84.9	84.1	83.3	87.4	88.4	87.2	26.8	29.6	27.3	23.0
<b>Retrieval-augmented Reasoning</b>												
Precedent-based-RAG-Qwen2.5-32B	82.9	85.9	87.2	86.2	82.1	85.7	86.2	85.4	22.3	31.1	21.1	18.2
Precedent-based-RAG-QwQ-32B	83.6	86.1	87.2	86.3	82.3	85.6	86.0	85.3	25.8	27.9	25.4	21.2
Precedent-based-RAG-Qwen2.5-72B	83.2	86.3	88.1	86.7	82.2	85.9	87.1	85.9	25.8	35.1	27.9	24.0
<b>Agentic Retrieval-augmented Reasoning</b>												
GLARE-Qwen2.5-32B(ours)	85.4	88.1	89.1	88.2	84.5	88.1	87.8	87.4	33.0	33.4	33.0	32.8
GLARE-QwQ-32B(ours)	<b>86.5</b>	<b>88.8</b>	<b>89.5</b>	<b>88.8</b>	<b>86.2</b>	<b>89.0</b>	<b>89.6</b>	<b>88.8</b>	<b>36.0</b>	<b>35.7</b>	<b>35.9</b>	<b>35.4</b>

Table 5: Performance comparison on the CMDL dataset. The best results are in bold.

**PLJP** utilize the domain models to provide candidate labels and find the proper precedents efficiently, and the large models will make the final prediction with an in-context precedents comprehension. We follow the topological dependencies among the three subtasks, employing BERT as the predictive model, SAILER as the retrieval model, and QwQ as the implementation of the LLM. Consistent with the original paper, we set the number of candidate labels to 3 and the number of retrieved precedents to 3. While PLJP considers only 42 common charges, our work covers all 164 charges, including both frequent and rare ones, thereby enabling a more comprehensive legal judgment prediction.

## F Results of CMDL dataset

The results are reported in Table 5. Our method demonstrated the best performance on all tasks and even outperformed the powerful DeepSeek-R1. Compared with direct reasoning and static RAG, GLARE simulates comparative reasoning by dynamically expanding the decision space to include confusing candidates, then retrieving exclusionary logic from precedents and statutes to identify the correct judgment.

## G Human Evaluation

To validate the overall effectiveness of the GLARE method, we conducted a human evaluation by scoring the judgment prediction processes generated by different approaches. Specifically, we hired three graduate students with strong legal expertise and extensive experience in the field. We randomly selected 50 cases and applied three methods(Direct Reasoning, Retrieval-augmented Reasoning, and

GLARE) to analyze and generate judgment predictions. After anonymizing the method names, we provided them with the fact descriptions of the cases, the correct judgment outcomes, and the predictions generated by three methods. They were then asked to score each result from 1 to 5 across the following three dimensions: (1) *Accuracy*: the model’s ability to correctly predict relevant legal provisions and charges; (2) *Interpretability*: the extent to which the reasoning logic and intermediate steps provided by the model are easily understandable by humans; and (3) *Trustworthiness*: the reliability of the prediction results.

As shown in Table 6, the GLARE method achieved the highest scores across all three dimensions, demonstrating that it not only accurately predicts legal judgments but also exhibits strong interpretability and trustworthiness. Its modular design and flexible invocation ensure that each step of the reasoning process is grounded in legal knowledge.

Method	Acc.	Expl.	Interp.	Avg.
Direct Reasoning	4.02	3.78	3.84	3.88
Precedent-based RAG	4.28	4.13	4.21	4.21
<b>GLARE(ours)</b>	<b>4.53</b>	<b>4.61</b>	<b>4.66</b>	<b>4.60</b>

Table 6: Human evaluation of reasoning performance, with scores from 1 to 5. The best results are in bold.

## H Analysis of Reasoning Enhancement

While GLARE improves performance by supplementing missing legal knowledge, we emphasize that its contribution goes beyond knowledge injection. GLARE enhances the intrinsic reasoning ability of LLMs by introducing a dynamic and structured process that guides legal decision-making.

We support this claim with three research questions and corresponding experiments.

**RQ1: Do LLMs already possess sufficient legal knowledge?** We first evaluate the legal knowledge coverage of LLMs by prompting different-sized models to generate applicable law articles. As shown in Table 7, although coverage improves with model size (from 42% for Qwen2.5-7B to 89.14% for Qwen2.5-72B), a notable portion of long-tail or specialized articles remains uncovered. This demonstrates the necessity of knowledge supplementation, especially in cases involving rare charges or legal concepts.

Model	Correct Number	Ratio
Qwen2.5-7B	147/350	42%
Qwen2.5-32B	292/350	83.43%
Qwen2.5-72B	312/350	89.14%

Table 7: Coverage of law articles generated by models of different sizes.

**RQ2: Is knowledge supplementation alone sufficient to enhance prediction?** To isolate the effect of pure knowledge injection, we provide all law articles and charges to the models as context. Interestingly, this strategy fails to improve and in some cases even degrades. As shown in Table 8, Qwen2.5-7B shows a decline in both charge and article prediction when all legal information is directly injected. This suggests that smaller models lack the capacity to utilize such information effectively. Thus, knowledge supplementation alone does not guarantee improved reasoning unless it is processed in a structured and selective manner.

Model	Charge		Law Article	
	Acc.	Ma-P	Acc.	Ma-F
Qwen2.5-7B	66.90	60.24	68.10	52.37
w/ All Law	62.90	57.15	58.90	45.56
QwQ-32B	81.57	77.25	82.35	73.81
w/ All Law	82.31	76.55	81.77	72.87
Qwen2.5-72B	74.20	69.77	76.25	64.38
w/ All Law	78.38	74.60	80.16	72.87

Table 8: Effect of injecting all law articles and charges into models of varying sizes. “w/ All Law” indicates that all legal knowledge is explicitly provided as input.

**RQ3: How does GLARE enhance reasoning ability?** GLARE is an agentic reasoning framework that dynamically integrates legal knowledge

throughout the decision-making process, improving discriminative reasoning capabilities.

- **CEM Module Expands Reasoning Breadth:** By expanding the average number of preliminary candidate charges from 1.33 to 9.19 as shown in the figure 5 (b), it prevents premature narrowing of the decision space that could lead to misjudgment, significantly broadening the coverage of the reasoning process.
- **LSAR Module Deepens Reasoning Depth:** The model autonomously detects knowledge gaps and retrieves relevant legal information when needed, thus avoiding interruptions in the reasoning chain caused by knowledge deficiencies.

**RQ4: Does GLARE perform genuine legal reasoning or merely surface-level pattern matching?** A key concern is whether the PRD module simply retrieves near-identical cases and imitates their outcomes, rather than learning transferable reasoning paths. To examine this question, we conduct two experiments: a Temporal Out-of-Distribution (OOD) Validation and a Logic-Perturbation Test.

**Temporal OOD Validation.** To evaluate generalization under distribution shift, we use the CJO22 dataset (Wu et al., 2023), which contains cases published after 2022 by the Supreme People’s Court of China. We use the training split of CAIL2018 (cases published before 2018) as the PRD retrieval library, and evaluate on 1,000 uniformly sampled cases from CJO22 used exclusively for testing. This strict temporal separation ensures that the test cases come from a distribution distinct from that used to construct the PRD library. As shown in Table 9, GLARE consistently outperforms both Direct Reasoning and Retrieval-augmented Reasoning across all three subtasks. These results suggest that the PRD module generalizes to temporally unseen cases rather than relying on memorized judgment trajectories.

**Logic-Perturbation Test.** To further test whether the model genuinely follows the reasoning paths demonstrated by precedents, we conduct a logic-perturbation experiment. Specifically, when constructing precedent reasoning paths, we randomly replace the correct charge with an incorrect but semantically similar charge, and then use Qwen2.5-72B to generate a plausible yet flawed reasoning path based on that incorrect label. We

Methods	Charge				Article				Prison Term			
	Acc.	Ma-P	Ma-R	Ma-F	Acc.	Ma-P	Ma-R	Ma-F	Acc.	Ma-P	Ma-R	Ma-F
<i>QwQ-32B</i>												
Direct Reasoning	81.6	85.0	78.3	79.0	83.6	80.6	73.3	74.5	18.5	29.5	19.6	16.3
Retrieval-augmented Reasoning	87.9	88.4	85.1	85.2	89.3	87.4	84.5	84.7	34.0	34.9	33.7	30.5
<b>GLARE (ours)</b>	<b>88.4</b>	<b>89.8</b>	<b>86.8</b>	<b>87.1</b>	<b>89.6</b>	<b>89.8</b>	<b>86.1</b>	<b>86.9</b>	<b>39.8</b>	<b>37.6</b>	<b>38.0</b>	<b>37.5</b>

Table 9: Temporal OOD validation on CJO22.

compare three variants against the full GLARE model:

- **w/ PRD-no-reasoning:** PRD includes only precedent facts and labels, without reasoning paths. This setting evaluates the importance of providing the reasoning path.
- **w/ PRD-logic-perturbation:** PRD includes reasoning paths deliberately constructed around an incorrect but similar charge. This setting tests whether the model truly follows the demonstrated reasoning logic.
- **w/o PRD:** The PRD module is removed entirely, assessing the overall contribution of precedent-based guidance.

## I Use of AI Assistants

We use ChatGPT to assist in polishing the language of this paper.

Methods	Charge		Article	
	Acc.	Ma-F	Acc.	Ma-F
<b>GLARE (ours)</b>	<b>89.7</b>	<b>88.6</b>	<b>91.3</b>	<b>88.5</b>
w/ PRD-no-reasoning	87.7	85.6	88.2	83.1
w/ PRD-logic-perturbation	77.9	75.5	79.6	72.4
w/o PRD	80.0	78.1	81.0	74.6

Table 10: Logic-perturbation analysis of the PRD module.

As shown in Table 10, we observe that performance under logic perturbation even drops below the w/o PRD setting. This indicates that the model follows the demonstrated but incorrect reasoning logic, suggesting that precedents function as demonstrations that actively guide the model’s reasoning process. Furthermore, when only the facts and labels of precedents are provided without reasoning paths, performance lies between the w/o PRD setting and our method. This demonstrates that complete precedent analysis including explicit reasoning is essential for guiding the model’s legal reasoning.

Overall, GLARE enables LLMs to simulate a more robust and legally grounded decision process.

## Instructions of Invoking External Modules

You are a reasoning assistant with the ability to call external tools. You have access to the following special tools:

**Expansion Tool:** Expands the initially predicted candidate charges. Write `<|begin_expand|>` Charge 1, Charge 2, ... `<|end_expand|>`.

The system will use existing tools to expand the charges you provide, obtaining a more comprehensive list of candidate charges while also supplying the reasoning paths from similar cases. The results will be returned in the format:

`<|begin_expand_result|>` Expanded charges, Reasoning paths from precedents  
`<|end_expand_result|>`.

You may use the expansion tool only once, and it must be called immediately after obtaining the initial predicted candidate charges.

**Search Tool:** When encountering multiple matching charges, performs a web search to clarify the distinctions between them. Write `<|begin_search_query|>` How to distinguish between [Charge A] and [Charge B] `<|end_search_query|>`.

Next, the system performs a search and analysis of web pages, outputting the information in the specified format:

`<|begin_search_result|>` ...Search results... `<|end_search_result|>`.

You may perform multiple searches, with a maximum attempt limit of {MAX\_SEARCH\_LIMIT}.

### Example:

Assistant:

I need to expand the charges of “Causing Death by Negligence,” “Endangering Public Security by Negligent Use of Dangerous Methods,” and “Gross Responsibility Accident Crime.”

Assistant:

`<|begin_expand|>` Causing Death by Negligence, Endangering Public Security by Negligent Use of Dangerous Methods, Gross Responsibility Accident Crime  
`<|end_expand|>`

(System returns expansion results)

`<|begin_expand_result|>` Expanded charges, Reasoning paths from similar cases  
`<|end_expand_result|>`

Assistant:

I need to identify the key differences between “Causing Death by Negligence” and “Gross Responsibility Accident Crime.”

Assistant:

`<|begin_search_query|>` Distinguishing between Causing Death by Negligence and Gross Responsibility Accident Crime `<|end_search_query|>`

(System returns search results)

`<|begin_search_result|>` ...Search results... `<|end_search_result|>`

Figure 7: The instruction used for GLARE.

### User prompts for GLARE

Your task involves predicting the applicable law articles and charges for a given case description. Based on the case facts, you should identify the most relevant law article and charge using your legal knowledge and external knowledge.

#### **Procedure:**

- Carefully analyze the provided case description.
- Generate all possible preliminary predictions.
- Use available tools to expand the list of candidate charges, obtaining a more comprehensive set of potential charges along with reasoning paths from precedents.
- Analyze each candidate charge's applicability to the current case using syllogistic reasoning, referencing the reasoning paths from similar cases.
- When multiple charges remain candidates, conduct web searches to clarify distinctions between them (e.g., "How to distinguish between Negligent Homicide and Gross Responsibility Accident Crime?").
- Continue the syllogistic analysis by combining search results with case facts. If search results contain unclear information, perform additional searches as needed (e.g., "What constitutes production operations in Gross Responsibility Accident Crime?"), ultimately determining the most appropriate legal provision and charge.

**Case description:** {fact}

**Defendant:** {criminal}

#### **Important Notes:**

All reasoning should be strictly grounded in the facts of the present case.

Figure 8: The prompt used for GLARE.