GLARE: Agentic Reasoning for Legal Judgment Prediction

Anonymous ACL submission

Abstract

Legal judgment prediction (LJP) has become increasingly important in the legal field. In this paper, we identify that existing large language models (LLMs) have significant problems of insufficient reasoning due to a lack of legal knowledge. Therefore, we introduce GLARE, an agentic legal reasoning framework that dynamically acquires key legal knowledge by invoking different modules, thereby improving the breadth and depth of reasoning. Experiments conducted on the real-world dataset verify the effectiveness of our method. Furthermore, the reasoning chain generated during the analysis process can increase interpretability and provide the possibility for practical applications.

1 Introduction

Legal judgment prediction (LJP) is an important task in legal natural language processing (NLP), aims to make correct judgment predictions based on the case's fact description (Liu et al., 2023). The judgment predictions include law articles, charges, and terms of penalty (Xu et al., 2024). 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 across 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., 2022). 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. In practice, they tend to predict the most likely charges without comparing them to similar alternatives, and their reasoning chains are often short and lacking in meaningful intermediate steps. These issues become especially clear in cases involving rare or confusing charges, where accurate judgment depends on subtle distinctions and careful reasoning. Although models may produce step-by-step outputs in such scenarios, the reasoning often stays at a surface level, focusing on pattern matching rather than legal principles.

We argue that the main reason for the limited performance of reasoning models in legal judgment tasks is not a lack of reasoning ability, but a lack of the specialized knowledge that legal reasoning depends on. Effective legal analysis requires long-tail legal knowledge, such as determining the applicability of specific statutes. In some cases, this knowledge is even absent from official legal texts. When such information is missing, models struggle to produce complete and trustworthy reasoning chains as shown in Figure 1. These observations highlight the need for domain-specific knowledge augmentation mechanisms that can dynamically supply essential information during the reasoning process.

To address the knowledge gaps in legal reasoning, we propose GLARE (*A**G**entic **L**eg**A**l **R**easoning Fram**E**work*), a modular system that enables language models to dynamically acquire key legal knowledge to improve the breadth and depth of reasoning. 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,

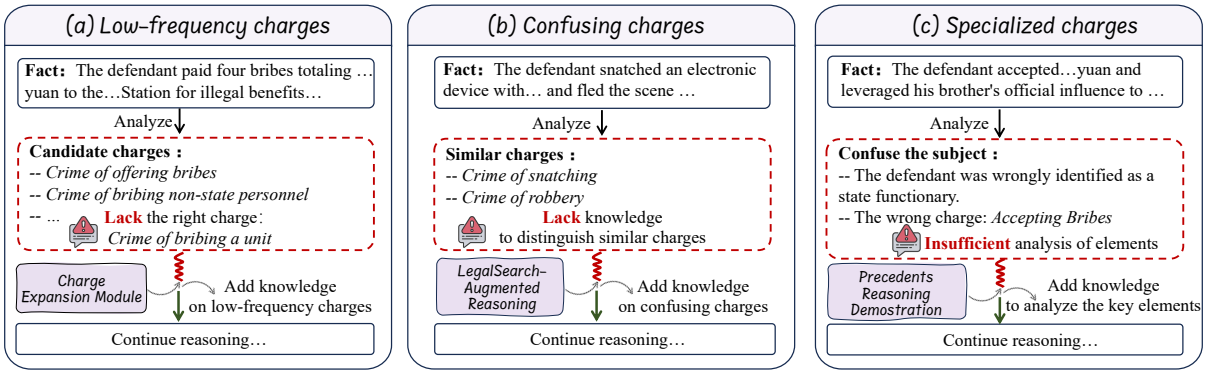

Figure 1: Lack of knowledge in three aspects: (a) Lack knowledge of low-frequency charges. (b) Lack knowledge of confusing charges. (c) Lack knowledge to analyze the key elements of the charges with strong professionalism.

the **Precedents Reasoning Demonstration** (PRD) module is built on reasoning paths that are constructed offline from real legal cases. During inference, the model retrieves the most relevant precedents through semantic search and learns from their reasoning chains via in-context learning. Finally, the **Legal Search-Augmented Reasoning** (LSAR) module allows the model to detect knowledge gaps and retrieve supporting legal information when needed. 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. Retrieved content is structured and injected into the reasoning process to support more accurate conclusions. By integrating essential legal knowledge, the model achieves more trustworthy and transparent judgment prediction.

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 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 has 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., 2024) to enhance LLMs by incorporating external legal knowledge. However, existing LLM-based methods struggle to utilize comprehensive legal knowledge (Fei et al., 2023) 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 and Yang, 2023) proposed legal syllogism reasoning to improve performance on LJP task. ADAPT (Deng et al., 2024b; Yuan et al., 2024) further established a comprehensive workflow for LJP that enables discriminative reasoning

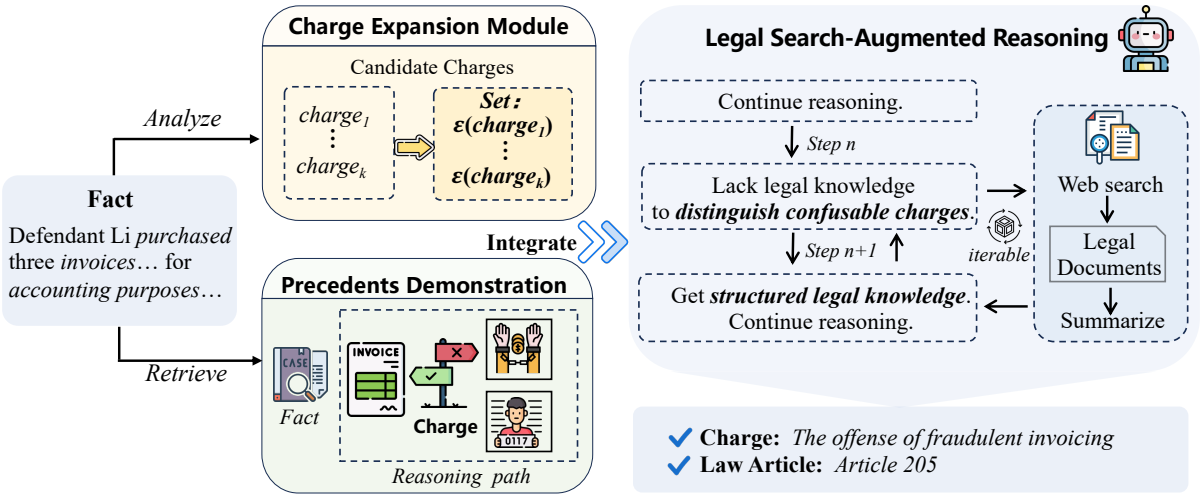

Figure 2: Overview of our agentic legal reasoning framework. LLMs can utilize three external modules to acquire knowledge: Charge Expand Module expands a diverse set of charges, precedents retrieved from offline built database can provide in-context learning, Legal Search-Augmented Reasoning allows the model to detect knowledge gaps and retrieve supporting legal information.

in LLMs. However, these approaches primarily rely on the LLMs 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 defendent. Following previous works (Shui et al., 2023; Wei et al., 2025), we exclude the task of sentencing prediction from our scope as its subjective nature brings challenges that are not well aligned with the current capabilities of large language models.

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 description 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:

$$(f, \mathcal{M}) \rightarrow (\mathcal{R}, p).$$

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 case facts and generating preliminary candidate charges. To prevent premature narrowing of the decision space, it 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 in detail.

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:

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

where top-$k_{\text{same}}(c)$ and 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 $\mathcal{E}_2(c)$.

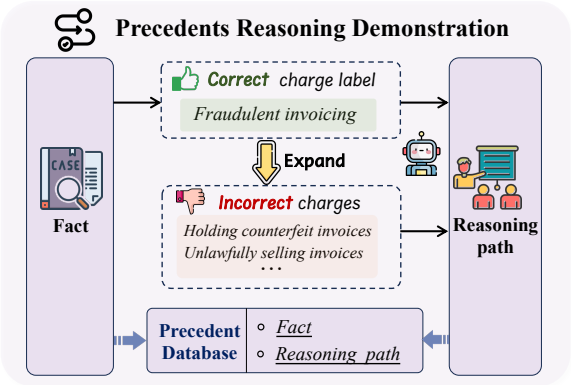

Figure 3: The module of Precedents Reasoning Demonstration: LLM analyzes the reasons for the selection or exclusion of each charge based on facts, thereby generating the reasoning path of precedents.

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:

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

3.4 Precedents Reasoning Demonstration

Previous precedent-based approaches (Wu et al., 2023; Chen and Zhang, 2023; 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.

To address this issue, we construct reasoning-augmented precedents that make the decision logic explicit. The precedent database is constructed using the CAIL2018 and CMDL training sets, comprising 163,563 cases that cover 190 charges and include both single-defendant and multi-defendant scenarios, which 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[1]. This reasoning is generated **offline** and stored together with the case facts.

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 *fact-specific questions*. The system retrieves relevant legal texts from the web in real time, enabling up-to-date and context-related augmentation.

We further ground the model's reasoning in a **syllogistic structure**: the retrieved legal context serves as the major premise, the case fact as the minor premise, and the conclusion is derived through logical alignment (Jiang and Yang, 2023; He et al., 2025). This structure helps the model remain grounded in factual evidence and reduce hallucinations. The overall reasoning process is formalized as an iterative function:

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

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.

[1]We provide the detailed prompt and examples for synthesizing reasoning paths in Appendix C

4 Experiments

4.1 Datasets and Evaluation

We conducted 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 sampled 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).

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 dataset.

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 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

Methods	Charge				Law Article			
	Acc.	Ma-P	Ma-R	Ma-F	Acc.	Ma-P	Ma-R	Ma-F
Classification Methods								
TopJudge	52.1	50.9	45.7	43.5	52.8	47.7	43.8	41.2
LADAN	76.7	73.4	71.0	69.5	77.5	71.0	69.2	67.5
NeurJudge	74.7	77.7	71.5	71.5	77.4	80.7	74.6	74.3
BERT	85.8	83.4	86.6	83.3	85.8	80.4	82.8	79.9
Lawformer	71.3	58.2	62.7	57.8	72.9	58.1	61.4	56.9
Direct Reasoning								
Qwen2.5-32B	74.5	75.3	69.3	69.1	77.1	73.3	66.6	67.1
QwQ-32B	82.5	86.9	80.5	80.9	84.0	83.1	76.1	77.0
Qwen2.5-72B	76.6	78.9	72.2	72.3	77.7	73.4	66.8	67.3
DeepSeek-R1-671B	84.8	86.3	81.3	81.7	87.2	86.8	81.8	82.6
Retrieval-augmented Reasoning								
Precedent-based-RAG-Qwen2.5-32B	88.5	88.2	85.8	85.7	89.4	87.2	83.7	84.5
Precedent-based-RAG-QwQ-32B	89.4	89.9	87.3	87.1	90.4	88.4	85.2	85.4
Precedent-based-RAG-Qwen2.5-72B	88.1	87.5	85.1	84.9	89.4	86.8	83.9	84.0
Search-o1-QwQ-32B	81.8	85.3	78.8	79.3	83.9	83.3	76.4	77.4
Agentic Retrieval-augmented Reasoning								
GLARE-Qwen2.5-32B(ours)	**89.8**	89.8	87.8	87.8	90.4	89.2	87.3	87.5
GLARE-QwQ-32B(ours)	89.7	**90.7**	**88.6**	**88.6**	**91.3**	**90.6**	**88.3**	**88.5**

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

few-shot settings (Brown et al., 2020). **Direct Reasoning** directly feeds the case facts into the LLM to predict the applicable law articles and charges, 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 (Guo et al., 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 (Yang et al., 2024), QwQ-32B (Team, 2025). *(2) Search-o1* (Li et al., 2025) dynamically retrieves external knowledge when it encounters uncertain or ambiguous knowledge in the general domain. We use reasoning model QwQ-32B (Team, 2025) in this setting.

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 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 [2] 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 D, and next we will analyze the experimental results:

1. Our method has demonstrated consistent performance improvements in both charge prediction and law article prediction tasks, highlighting the effectiveness of our agentic reasoning approach for LJP. Compared to the Direct Reasoning setting, our method improves charge prediction by 7.7% and law article prediction by 11.5% in F1 score. When compared with Retrieval-augmented Reasoning, it achieves an improvement of 1.5% on charge prediction and 3.1% on law article prediction in F1 score. In addition, our method not only performs well on the large reasoning models, but also effectively promotes the reasoning ability of the instruct models, indicating that our three modules effectively supplement legal knowledge and thereby enhance reasoning performance[3].

2. In contrast to direct reasoning, precedent-based RAG enhances prediction performance through precedent retrieval. Large reasoning

[2] https://serper.dev

[3] Appendix F discusses how our method enhances LLM reasoning beyond knowledge supplementation.

Methods	CAIL2018				CMDL			
	Charge		Law Article		Charge		Law Article	
	Acc.	Ma-F	Acc.	Ma-F	Acc.	Ma-F	Acc.	Ma-F
Direct Reasoning								
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
Retrieval-augmented Reasoning								
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
Agentic Retrieval-augmented Reasoning								
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)	**90.7**	**75.7**	**91.1**	**75.4**	**76.0**	**79.5**	**74.0**	**76.7**

Table 3: Performance comparison on difficult charges.

models like QwQ-32B and DeepSeek-R1-671B outperform other instruct models in direct reasoning, indicating that LJP inherently requires multi-step reasoning and slow thinking. Precedent-based RAG improves performance across models of various scales by incorporating precedent retrieval. For example, QwQ-32B sees an 8.39% F1 improvement in law article prediction. However, precedent-based RAG only provides the case facts and labels of precedents, leading models to rely on similarity matching and copy judgment predictions rather than truly reason. Additionally, Search-o1 retrieves case facts which may introduce noises, rather than specific legal knowledge, thus underperforming compared to direct reasoning.

3. LLM-based methods outperform classification methods. However, BERT outperforms LLMs like Qwen2.5-72B-Instruct in direct reasoning. The key reasons are: (1) BERT frames predictions as multi-class classification tasks, enabling direct mapping from facts to fixed labels, which aligns well with the task. In contrast, LLMs take a generative approach and without legal-specific training they often fail to make accurate predictions. (2) Our dataset includes many rare and confusing charges. Fine-tuned BERT models trained on legal corpora can better distinguish these nuanced charges, while even large LLMs lack the domain knowledge needed for such difficult charges.

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 a diverse set of candidate charges. **(2) w/o PRD:** The Precedents Reasoning Demonstration module is removed, so the model cannot refer to 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.

Methods	Charge		Law Article	
	Acc.	Ma-P	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)	**89.7**	**88.6**	**91.3**	**88.5**

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

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 crucial role of precedent reasoning path in enhancing legal judgment prediction. Removing CEM weakens the model's ability to recognize ambiguous or low-frequency charges, while LSAR helps the model fill knowledge gaps by retrieving authoritative legal information. Overall, the GLARE framework performs best across all metrics.

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*

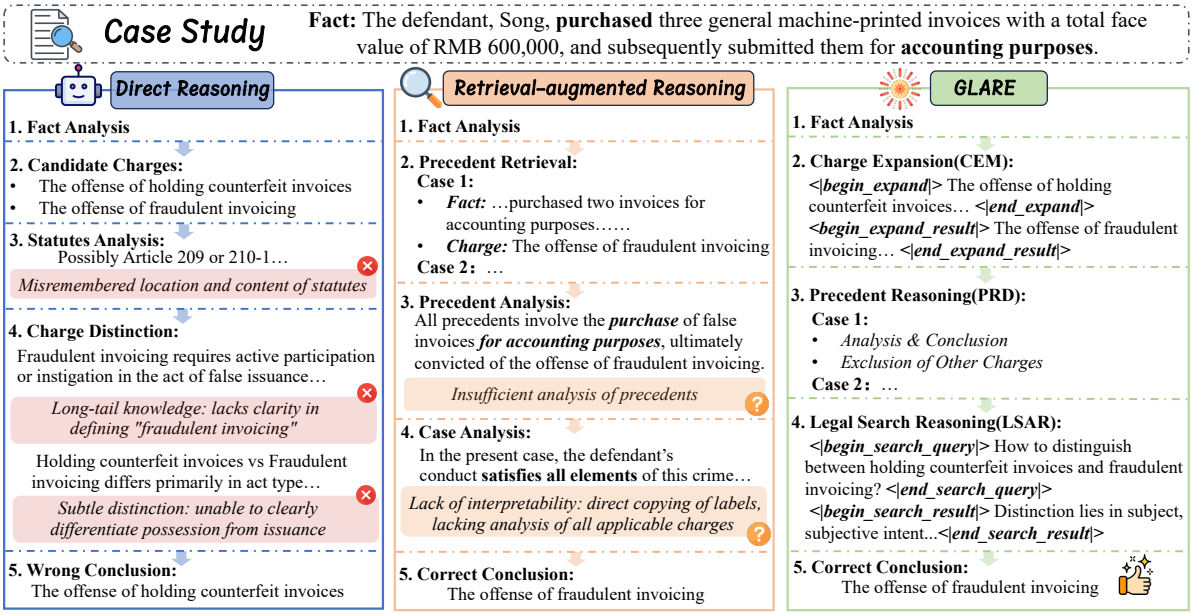

Figure 4: Case Study. The red part highlights the model's limitations due to insufficient internal knowledge, while the yellow part demonstrates the lack of interpretability in vanilla precedent-based RAG reasoning.

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 the Figure 4, we conducted case study on three LLM-based methods to further verify the effectiveness and interpretability of our method. Direct Reasoning relies on the LLM's internal knowledge, which may be inaccurate or insufficient, leading to incorrect judgments. RAG methods often lack explicit links between retrieved cases and final decisions, making it hard to trace how external knowledge affects reasoning. However, our method ensures that each reasoning step has a clear knowledge basis through the explicit invocation of three modules, thereby improving interpretability. We also conducted a human evaluation of the quality of the model-generated replies in Appendix E.

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 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.

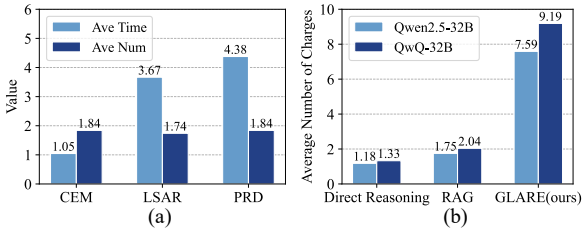

Figure 5: (a) Efficiency analysis of each module. (b) Average charge numbers of different methods.

5 Conclusion

In this study, we propose a novel framework, GLARE, to address the legal gaps in legal reasoning. GLARE dynamically acquires key legal knowledge to improve the breadth and depth of reasoning. Experimental results demonstrate the effectiveness of our approach, which not only improves prediction performance but also generates complete reasoning chains that enhance the interpretability of LJP tasks. 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 the 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 noises in the process of generating rich candidate charges, our experiments show that the influence of these noises 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.

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.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv: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 X Zhang. 2023. Case law grounding: Using precedents to align decision-making for humans and ai. *arXiv preprint arXiv:2310.07019*.

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. *Preprint*, arXiv:2507.08806.

Chenlong Deng, Kelong Mao, and Zhicheng Dou. 2024a. Learning interpretable legal case retrieval via knowledge-guided case reformulation. *arXiv preprint arXiv:2406.19760*.

Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. 2024b. Enabling discriminative reasoning in llms for legal judgment prediction. *arXiv preprint arXiv:2407.01964*.

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, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

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. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

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

Congqing He, Tien-Ping Tan, Sheng Xue, and Yanyu Tan. 2025. Simulating judicial trial logic: Dual residual cross-attention learning for predicting legal judgment in long documents. *Expert Systems with Applications*, 261:125462.

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, and 1 others. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

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. *arXiv preprint arXiv:2501.05366*.

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. *arXiv preprint arXiv:1707.09168*.

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. *Preprint*, arXiv:2312.05762.

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. *arXiv preprint arXiv: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. *arXiv preprint arXiv:2402.07927*.

TYSS Santosh, Mohamed Hesham Elganayni, Stanisław Sójka, and Matthias Grabmair. 2024. Incorporating precedents for legal judgement prediction on european court of human rights cases. *arXiv preprint arXiv:2409.18644*.

Jeffrey A Segal. 1984. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. *arXiv preprint arXiv:2310.11761*.

Olga Shulayeva, Advaith 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. *arXiv preprint arXiv:1710.09306*.

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. *arXiv preprint arXiv:2401.06805*.

Bin Wei, Yaoyao Yu, Leilei Gan, and Fei Wu. 2025. An llms-based neuro-symbolic legal judgment prediction framework for civil cases. *Artificial Intelligence and Law*, pages 1–35.

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. *arXiv preprint arXiv:2310.09241*.

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. *Preprint*, arXiv:1807.02478.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, 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. *arXiv preprint arXiv:2004.02557*.

Nuo Xu, Pinghui Wang, Junzhou Zhao, Feiyang Sun, Lin Lan, Jing Tao, Li Pan, and Xiaohong Guan. 2024. Distinguish confusion in legal judgment prediction via revised relation knowledge. *Preprint*, arXiv:2408.09422.

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. *Preprint*, arXiv:2410.02507.

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.

Jianqiiu 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. *arXiv preprint arXiv:2403.03288*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

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 of 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 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 Prompts of Precedents Reasoning Demonstration

We construct reasoning-augmented precedents that make the decision logic explicit. We first expand the original charge label into a set of similar charges. Then given the case fact f, the correct charge c, and the set of alternatives C, we prompt LLM to generate a reasoning path r. Detailed prompts are shown in Figure 6. Next, we present a specific example of 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.*

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.

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 was not merely possessing counterfeit invoices but fabricating 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 Implicated 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 directly defraud.* ***(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*

Methods	Charge				Law Article			
	Acc.	Ma-P	Ma-R	Ma-F	Acc.	Ma-P	Ma-R	Ma-F
Direct Reasoning								
Qwen2.5-32B	70.4	76.0	80.2	77.2	73.5	79.3	81.0	79.3
QwQ-32B	77.9	81.7	83.7	82.1	79.5	83.4	85.1	83.5
Qwen2.5-72B	73.6	77.5	80.6	78.5	76.1	80.1	80.9	79.8
DeepSeek-R1-671B	81.4	84.1	84.9	84.1	83.3	87.4	88.4	87.2
Retrieval-augmented Reasoning								
Precedent-based-RAG-Qwen2.5-32B	82.9	85.9	87.2	86.2	82.1	85.7	86.2	85.4
Precedent-based-RAG-QwQ-32B	83.6	86.1	87.2	86.3	82.3	85.6	86.0	85.3
Precedent-based-RAG-Qwen2.5-72B	83.2	86.3	88.1	86.7	82.2	85.9	87.1	85.9
Agentic Retrieval-augmented Reasoning								
GLARE-Qwen2.5-32B(ours)	85.4	88.1	89.1	88.2	84.5	88.1	87.8	87.4
GLARE-QwQ-32B(ours)	**86.5**	**88.8**	**89.5**	**88.8**	**86.2**	**89.0**	**89.6**	**88.8**

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

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 Results of CMDL dataset

The results are reported in Table 5. Our method demonstrated the best performance on both tasks and even outperformed the powerful DeepSeek-R1. Compared with direct reasoning and static RAG, GLARE can autonomously and dynamically supplement knowledge, thereby extending the reasoning chain and enhancing interpretability.

E Human Evaluation

To validate the interpretability 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 the 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.

F 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.

Method	Accuracy	Explainability	Trustworthiness	Average
Direct Reasoning	4.02	3.78	3.84	3.88
Retrieval-augmented Reasoning	4.28	4.13	4.21	4.21
GLARE(ours)	**4.53**	**4.61**	**4.66**	**4.60**

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

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 pipeline, improving both the breadth and depth of reasoning.

- **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.

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.