Comprehensive is not perfect: Enhancing LLMs with Expert Notes

Anonymous ACL submission

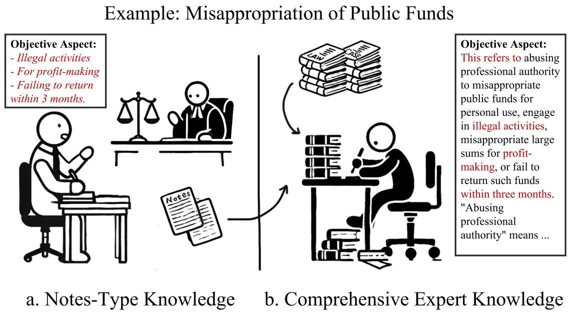
Abstract

Large Language Models (LLMs) are increasingly employed in specialized fields, such as the legal domain, where expert knowledge is essential to overcome their inherent limitations. However, acquiring comprehensive expert knowledge is often costly and impractical. To mitigate the reliance, researchers have explored leveraging fragmented expert insights to help LLMs mimic expert reasoning. However, such approaches often lack the practical experience that expert provides. In this paper, we introduce a novel form of expert experience: Notes-type Knowledge. It is less formalized and precise but is more accessible and contains the practical expertise often missing in LLMs. Focusing on the Four-element Theory (FET) in Chinese criminal law, we annotate the four elements knowledge in notes-type for 194 charges, and purpose a Notes-guided LLM method to integrate LLMs with notes-type knowledge. Experiments on Similar Charge Disambiguation and Legal Case Retrieval tasks show that the approach outperforms LLMs and achieves performance comparable to that with comprehensive expert knowledge.

1 Introduction

When applying LLMs to a specific domain, such as the legal fields, it is often necessary to incorporate expert knowledge to supplement the ability deficiency of general LLMs (Zhou et al., 2024; Cui et al., 2023; Yao et al., 2024). This knowledge, meticulously curated by experts, ensures accuracy and completeness (Li et al., 2024a; Cheong et al., 2024), but is costly and impractical to acquire for new tasks.

To reduce reliance on comprehensive expert knowledge, researchers have explored leveraging fragmented and easily obtainable expert insights to guide LLMs in mimicking expert reasoning processes. For example, in legal charge prediction, Jiang et al.(Jiang and Yang, 2023) proposed reasoning based on the legal syllogism, while others(Yuan et al., 2024; Deng et al., 2023) utilized the Four-element Theory, both of which are widely accepted frameworks in Chinese legal practice. These approaches have demonstrated superior performance compared to general Chain-Of-Thought (CoT) reasoning(Kojima et al., 2022) in legal tasks.


Example: Misappropriation of Public Funds
Objective Aspect:
- Illegal activities
- For profit-making
- Failing to return within 3 months.
Notes
LAW
CASE
Objective Aspect:
This refers to abusing professional authority to misappropriate public funds for personal use, engage in illegal activities, misappropriate large sums for profit-making, or fail to return such funds within three months. "Abusing professional authority" means ...

a. Notes-Type Knowledge b. Comprehensive Expert Knowledge

Figure 1: Notes-type Knowledge and Comprehensive Expert Knowledge

Despite these advancements, it remains unclear whether mimicking expert reasoning can truly replace comprehensive expert knowledge or merely serves as a practical compromise in its absence. In this work, we aim to explore two questions: (1) Can LLMs achieve performance comparable to using comprehensive expert knowledge by imitating the reasoning process of legal experts? (2) If not, can we identify a cost-effective alternative that leverages expert insights without incurring the high costs of fine-grained annotation?

We focus on the Four-element Theory (FET) in Chinese criminal law, which delineates four essential components for establishing a criminal charge: Subject, Object, Subjective aspect, and Objective aspect. We compare the four elements derived from LLMs' internal knowledge with those based on carefully curated expert knowledge. Through human evaluation, we find that although the LLM-generated four elements are relatively accurate and standardized, they lack an understanding of the connections between charges and representative case plots. This motivates us to introduce more representative knowledge to complement the LLM itself.

Specifically, we introduce the ***Notes-type Knowledge***. We annotate the four elements in this forms for 194 charges in Chinese criminal law. Figure 1 shows the difference between comprehensive expert knowledge and notes-type knowledge. Notes-type knowledge resembles the notes a lawyer takes while working on a specific case, including key legal concepts and representative examples(e.g. typical action, plot, or result). As a by-product or intermediate manuscript, it is more accessible than comprehensive expert knowledge. Although it scored low on Completeness and Standardization, it has high Representativeness, which can complement the inherent knowledge of LLMs. Building on this, we propose a notes-guided method that combines the strengths of LLMs and notes-type knowledge.

We evaluate our approach on two legal tasks: Similar Charge Disambiguation (SCD) and Legal Case Retrieval (LCR). Experiments on the public SCD dataset (Liu et al., 2021) show that LLM-generated four-element reasoning trails expert-curated knowledge by 0.7 points in average F1-score, indicating that emulating experts does not yet match authentic expert knowledge. However, incorporating Notes-type Knowledge not only bridges this gap but also surpasses comprehensive expert knowledge, achieving an additional 0.54-point improvement in average F1-score and a 0.57-point improvement in average accuracy. This demonstrates that notes-type knowledge provides practical insights and highlights key statutory information, while LLMs effectively refine these inputs to enhance the overall quality of generated legal elements. We further evaluate the proposed method in the LCR task and find that integrating notes-guided knowledge helps the model extract more accurate details for case-specific elements from factual descriptions. On the public LeCaRDv2 dataset(Li et al., 2024c), notes-guided approach improves the model performance by an average of 12.66% in MRR.

Our contributions are as follows:

(1) We empirically demonstrate the limitations of emulating expert reasoning with LLM in legal tasks compared to using comprehensive expert knowledge.

(2) We introduce a novel form of expert experience, notes-type knowledge, which, while less structured and comprehensive than traditional expert-curated knowledge, is significantly easier to obtain and captures the nuances absent in LLMs. We annotate the four elements in notes-type and comprehensive expert knowledge for 194 charges in Chinese criminal law.

(3) We propose a notes-guided framework that integrates LLMs with notes-type knowledge, achieving results comparable to comprehensive expert knowledge in legal tasks while reducing annotation and construction costs.

2 Preliminary

2.1 Four Element Crime Composition Theory

In this paper, we adopt the Four-Element Crime Composition Theory (FET) to study how expert-driven and LLM-driven knowledge can complement each other to enhance legal reasoning and task performance. FET is one of the most widely recognized criminal theory in Chinese judicial practice (Liang, 2017). It specifies four essential elements that must be satisfied to establish criminal liability: **Subject, Object, Subjective aspect, and Objective aspect**. For example, the four elements of the Crime of Affray are as follows:

(1) Subject: Principal organizers and other active participants who have reached the age of criminal responsibility. (2) Object: Public order. (3) Objective Aspect: The act of assembling brawl, engaging in a brawl, resulting in the following consequences of serious injury. (4) Subjective Aspect: Direct intent, where the person knowingly and willfully engages in organizing or participating in the act of assembling brawl.

Four-element theory are widely incorporated in the legal AI domain to assist models in solving legal problems. For example, (Yuan et al., 2024) employ an auto-planning strategy to decompose legal rules into four aspects aligning with FET, (Deng et al., 2023) leverage model-generated four elements as minor premises in legal judgment analysis. However, most of the previous efforts depend on the LLM's internal knowledge. Whether LLMs understand the Four-element theory correctly has not been evaluated.

2.2 Four Elements from Experts

In our study, we explore two types of expert knowledge: Comprehensive expert knowledge and Notes-type knowledge. We annotate the four elements in both forms for 194 charges in Chinese criminal law. The annotation details are as follows:

Notes-Type Knowledge: Notes-type knowledge was annotated by four postgraduate law students. During annotation, they referred to criminal law articles and real judicial cases. They were instructed to focus on two aspects for each element: (1) Formalized Keywords, where they identified key terms for each element with relaxed constraints, allowing for practical interpretations. For instance, the object element can be concrete objects instead of abstract legal concepts, as shown in Table 6; (2) Common Scenarios, where they summarized typical situations based on case analysis and their understanding, such as recognizing "assault" as a typical situations of subjective aspect in the crime of intentional injury. This approach simulates the aggregation of informal notes from various cases handled by lawyers.

Comprehensive Expert Knowledge: Expert knowledge is finely annotated based on notes by a doctor of law. The annotations were based on criminal law articles, textbooks used in the Bar Examination and Juris Master Examination, as well as real judicial cases. The expert emulated the analytical process of lawyers by reviewing relevant laws, identifying key terms, and providing comprehensive explanations. This simulates the lawyer's detailed annotation process.

Drawing from previous work(Cui et al., 2024; Zhou et al., 2023), we define LLM-generated knowledge as information produced by a large language model based on its pre-trained knowledge and contextual prompts.

LLM-Generated Knowledge: We provide the model with legal articles and the definition of four elements in FET, prompting it to generate the four elements within this framework. The LLM is expected to autonomously identify and generate the four elements based on its learned understanding of legal concepts.

3 Does LLM handle FET Correctly?

In order to evaluate whether the LLM have already handle the Four-element Theory, we invite legal experts to compare the four elements generated by the LLM with two types of expert knowledge.

Methods	Precision	Completeness	Representativeness	Standardization
Notes-type	3.62	3.27	3.88	3.23
Comprehensive Expert	**4.69**	**4.65**	**4.48**	4.56
LLM-generated	4.12	3.79	3.60	4.33
Notes-guided	4.46	4.35	4.35	**4.69**
$\delta_{Notes\text{-}type}$	0.84	1.08	0.47	1.46
$\delta_{LLM\text{-}generated}$	0.34	0.56	0.75	0.36

Table 1: Performance Comparison of Four Elements Across Methods. δ denotes the score difference between the Notes-guided method and others. The Notes-guided method shows improvements across all dimensions, excelling in Representativeness over LLM-generated four elements and in Precision and Standardization over Notes-type four elements.

3.1 Human Evaluation

We evaluate the four elements produced by each approach from 4 dimensions: **Precision, Completeness, Representativeness, and Standardization**:

- Precision: Whether the key components are accurately identified.
- Completeness: Whether all necessary elements of the four-element theory are included.
- Representativeness: Whether the annotations highlight the most important legal elements and case details.
- Standardization: Whether the annotations are clear, consistent, and adhere to established norms for easy interpretation.

Each dimension was scored by two types of experts: one group with a pure legal background and another group with a combined background in law and Artificial Intelligence, all of whom have passed the bar examination. The experts were selected to balance domain expertise and interdisciplinary perspectives. Scores were averaged across the two groups. Details about 1-5 scale criteria and annotator background are provided in Appendix B.

3.2 Result

The results are shown in Table 1. The human evaluation results reveal that while both Notes-type four elements and LLM-generated four elements underperform compared to comprehensive expert annotations, their reasons differ.

Due to the lack of details, the Note-type four elements scored low on Completeness and Standardization. But focusing on key legal terms allows them to capture the most critical aspects of the charge, maintaining relatively well performance in Representativeness. In contrast, LLMs excel in Precision and Standardization due to their focus on the literal decomposition and restatement of legal provisions but fall short in fully explaining or

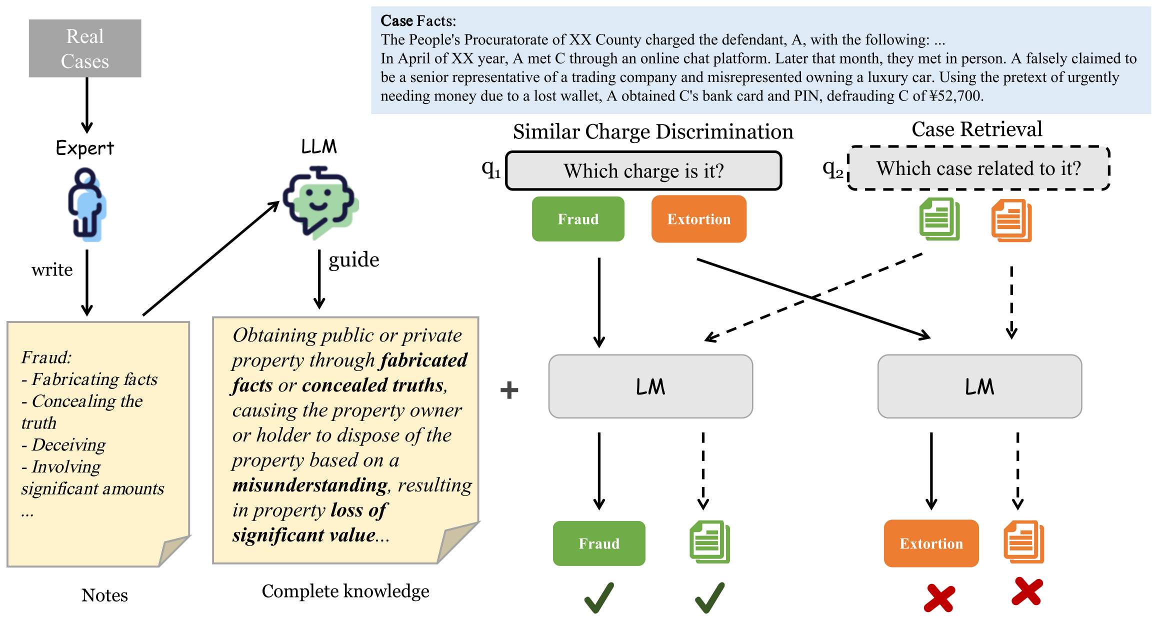

Real Cases
Expert
write
LLM
guide
Fraud:
- Fabricating facts
- Concealing the truth
- Deceiving
- Involving significant amounts
...
Notes
Obtaining public or private property through fabricated facts or concealed truths, causing the property owner or holder to dispose of the property based on a misunderstanding, resulting in property loss of significant value...
Complete knowledge
+
Case Facts:
The People's Procuratorate of XX County charged the defendant, A, with the following: ...
In April of XX year, A met C through an online chat platform. Later that month, they met in person. A falsely claimed to be a senior representative of a trading company and misrepresented owning a luxury car. Using the pretext of urgently needing money due to a lost wallet, A obtained C's bank card and PIN, defrauding C of ¥52,700.
Similar Charge Discrimination
q1 Which charge is it?
Fraud
Extortion
Case Retrieval
q2 Which case related to it?
LM
LM
Fraud
Extortion

Figure 2: Notes-guided FET Method.

analyzing the underlying legal concepts.

4 Notes-guided FET Method

To combine the organizational efficiency of LLMs with the representational precision of notes-style knowledge, we propose the Notes-guided approach, which leverages expert notes to guide LLMs in generating the four elements. As shown in Table 1, this hybrid method improves human evaluation performance across all dimensions. Compared to the four elements generated solely by LLMs, the Notes-guided approach achieves a 20.83% relative improvement in Representativeness. Compared to the original notes, it improves 33.03% in Completeness and 45.20% in Standardization, demonstrating the ability of the Notes-guided method to generate higher-quality four elements.

To further explore the performance of notes-guided method, we developed the Notes guide LLM framework for legal AI tasks with two questions: (1) Is notes-guided comparable with expert-guided? (2) Can notes-guided knowledge help more downstream tasks?

For the first question, we chose the Similar Charge Disambiguation (SCD) task, a challenging subset of charge prediction. This task requires the LLM to differentiate between similar charges based on factual descriptions and legal rules, demanding a deep understanding of the structural composition of criminal charges(Yuan et al., 2024; Li et al., 2024a). As shown in fig 2, in SCD task, we utilized the notes-guided four elements corresponding to the analyzed crimes as input guidance for the model. This allowed for a direct comparison between the notes-guided and expert-guided methods.

For the second question, we evaluate the utility of notes-guided knowledge in the Legal Case Retrieval (LCR) task, which involves retrieving relevant legal cases based on given case facts. It is an important step in the practice of analyzing cases and making judgments, and it requires the precise application of the four-element theory to interpret case facts. Additionally, the large-scale search pool in LCR renders expert annotations impractical, highlighting the value of a cost-effective and scalable annotation tool like notes. As shown in fig 2, in LCR task, we use expert notes to guide the LLM in generating case-specific four-element analyses.

5 Is Notes-Guided Comparable with Expert Carefully Curated Knowledge?

In the preceding section, the human evaluation demonstrated that combining notes with the LLM's internal knowledge has the potential to generate higher-quality four elements. In this section, we conduct a quantitative analysis using the Similar Charge Disambiguation task to investigate whether notes-guided knowledge can enhance the model's understanding of the Four-Element Theory and how it differs from expert-curated knowledge.

Prompt:
You are a lawyer specializing in criminal law. Based on Chinese criminal law,
please determine which of the following candidate charges the given facts align with.
The candidate charges and their corresponding four elements are as follows:
[Four Elements of Candidate Charges].
The four elements represent the four core factors of a charge.*[The basic concepts of the Four-element theory]*
Compare the case facts to determine which charge's four elements they align with,
thereby identifying the charge.

Table 2: Example of notes-guided instruction for charge disambiguation.

5.1 Method

To employ notes-guided knowledge for charge disambiguation, we adopt a unified process. As shown in Table 2. For each group of similar charges, the corresponding four elements generated through different methods were incorporated into the instructions. The LLM then used these enriched inputs to match the given case facts with the appropriate charge.

To align with the human evaluation in Section 3.1, we compared notes-guided four-elements with three other methods for gaining four-elements: Comprehensive Expert Knowledge, Notes-Type Knowledge, and LLM-Generated Knowledge. While the instructions followed the same format, the *[Four Elements of candidate charges]* were replaced with those derived from each respective method.

All four methods are represented as follows:

FET$_{\textbf{Expert}}$: Directly using expert-annotated four-element corresponding to the charges being analyzed.

FET$_{\textbf{Notes}}$: Directly using four-element notes corresponding for the charges being analyzed.

FET$_{\textbf{LLM}}$: Directly using the four elements of crimes generated by the LLM for the charges being analyzed.

FET$_{\textbf{Notes\_guided}}$: Using notes to guide the LLM in generating four elements for the charges being analyzed.

5.2 Dataset

We chose the dataset released by (Liu et al., 2021), which includes five charge sets with the largest number of cases. To evaluate performance on representative tasks, we selected three 2-label classification groups commonly examined in other datasets (Yuan et al., 2024): Fraud & Extortion (F&E), Embezzlement & Misappropriation of Public Funds (E&MPF), and Abuse of Power & Dereliction of Duty (AP&DD).

Charge Sets	Charges	Cases
F&E	Fraud & Extortion	3536 / 2149
E&MPF	Embezzlement & Misappropriation of Public Funds	2391 / 1998
AP&DD	Abuse of Power & Dereliction of Duty	1950 / 1938

Table 3: Distribution of charges in the GCI dataset. Cases denotes the number of cases in each category. Following (Liu et al., 2021), for a case with both confusable charges, the prediction of any one of the charges is considered correct.

The details of the classification groups are shown in Table 3. Following previous work (Liu et al., 2021; Yuan et al., 2024), we use Average Accuracy (Acc) and macro-F1 (F1) as evaluation metrics.

5.3 Baselines

To compare the performance of traditional methods and LLMs on SCD tasks, we evaluate several baseline models commonly used in previous work, including: GCI, CausalChain (Liu et al., 2021), Bi-LSTM (Zhou et al., 2016), Bi-LSTM+Att, and Bi-LSTM+Att+Cons.

To explore the effectiveness of notes-guided four elements in LLMs, we further consider other methods that introduced the Four-element theory into LLMs, including: GPT-4o(Achiam et al., 2023), GPT-4o$_{\text{Law}}$, Legal-COT(a variant of COT (Kojima et al., 2022)).

Details of each baseline is shown in Appendix C. For traditional models, we split the training and test sets 1:1. For LLMs, we take a zero-shot setting.

5.4 Results

As shown in Table 4, the FET$_{\text{Notes\_guided}}$ achieves the highest overall performance, surpassing the expert-guided method by 0.57 in average accuracy and 0.54 in average F1. This demonstrates that the guided method effectively combines the strengths

Model	**F&E**		**E&MPF**		**AP&DD**		**Average**	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LSTM	90.04	89.06	75.59	75.46	69.65	69.62	80.21	77.89
Bi-LSTM	90.43	89.83	76.08	75.78	71.12	70.50	80.65	78.21
GCI	90.41	89.14	89.01	88.63	81.01	80.90	84.49	82.35
CausalChain	90.45	89.21	81.25	80.09	80.03	79.89	82.76	80.16
Bi-LSTM+Att	91.56	91.05	82.29	82.11	73.70	73.65	83.41	81.27
Bi-LSTM+Att+Cons	92.05	91.55	83.02	82.69	80.72	80.64	85.36	83.42
GPT-4o	94.36	95.81	86.49	89.76	85.54	87.12	88.72	90.07
GPT-4o$_{Law}$	95.34	96.30	92.64	93.03	88.30	89.33	92.09	92.89
Legal-COT	94.99	96.27	90.50	90.99	87.81	88.14	89.95	90.85
FET$_{Notes}$	95.80	96.51	91.18	91.22	90.59	90.71	92.52	92.81
FET$_{LLM}$	95.73	96.56	91.87	92.01	89.61	89.69	92.40	92.75
FET$_{Expert}$	**96.06**	**96.69**	92.57	93.05	90.53	90.62	93.05	93.45
FET$_{Notes\_guided}$	95.97	96.65	**93.10**	**93.44**	**91.80**	**91.89**	**93.62**	**93.99**

Table 4: Results of Charge Disambiguation. FET means the Four-element theory framework with knowledge obtained from experts, experts' notes, LLM, and notes-guided LLM method. Highest results are in bold.

of expert notes and LLM knowledge. Below are some key findings:

Effectiveness of LLM's Internal Knowledge: The classification accuracy and F1 scores of LLM-based methods consistently surpass those of traditional models, indicating that the internal knowledge of LLMs is beneficial for domain-specific tasks, confirming the advantages of leveraging pre-trained knowledge from LLMs.

Importance of Human Guidance: Comparing GPT-4o, GPT-4o$_{\text{Law}}$, and Legal-CoT, the latter two methods outperform GPT-4o, highlighting the value of incorporating legal knowledge. However, both GPT-4o$_{\text{Law}}$ and Legal-CoT are outperformed by FET-based methods. This suggests that merely providing a theoretical framework (as Legal-CoT does) or supplying law articles without explicit legal theory guidance is insufficient for making legal decisions, as they rely solely on the LLM's internal knowledge.

FET$_{\text{Notes}}$ vs. FET$_{\text{LLM}}$ vs. FET$_{\text{Expert}}$: Among these methods, FET$_{\text{Expert}}$ achieves the best performance, aligning with human evaluations in Table 1 and validating the importance of detailed expert knowledge. As expected, FET$_{\text{Notes}}$ performs worse, with a reduction of 0.53 in average accuracy and 0.61 in average F1 compared to FET$_{\text{Expert}}$, which can be attributed to its less detailed annotations.

Complementarity Between Experts' Notes and LLM: Although the notes-type four elements scores lower than LLM-generated elements in human evaluation, they slightly outperform FET$_{\text{LLM}}$ in the SCD task, reflecting the value of expert-derived insights for charge determination. By refining the information in the notes through LLMs, the FET$_{\text{Notes\_guided}}$ method achieves the best performance, even surpassing FET$_{\text{Expert}}$. This result shows LLMs effectively leverage the additional information provided by notes, with their complementary strengths leading to optimal outcomes.

6 Can Notes-Guided Knowledge Benefit More Downstream Tasks?

In the previous section, we demonstrated that notes-guided knowledge enhances the LLM's comprehension of the Four-element theory. In this section, we use another prevalent scenario in legal practice, Legal Case Retrieval (LCR), to evaluate the potential of further applying notes-guided method in downstream legal tasks.

6.1 Method

To investigate the effect of using notes to guide knowledge for LCR, we propose three methods that progressively increase in their use of external knowledge:

BGE: It's a basic method that directly matches the query and candidate based solely on their case facts, without incorporating legal theories. We selected BGE-m3(Chen et al., 2023), a widely used embedding model for dense retrieval, due to its effectiveness in capturing semantic similarities in large-scale datasets.

FET$_{\text{LLM}}$: To facilitate comparison, we propose a method that integrates the Four-element theory into the retrieval process. For each case in the

Model	NDCG@10	NDCG@20	NDCG@30	R@1	R@5	R@10	R@20	R@30	MRR
QL	0.4438	0.4965	0.5372	0.0977	0.2831	0.4158	0.5517	0.6421	0.1969
BM25	0.4046	0.4650	0.5095	0.0681	0.2608	0.3889	0.5384	0.6467	0.1719
BERT	0.1511	0.1794	0.1978	0.0199	0.0753	0.1299	0.2157	0.2579	0.1136
Legal-BERT	0.1300	0.1487	0.1649	0.0186	0.0542	0.1309	0.1822	0.2172	0.0573
Lawformer	0.2684	0.3049	0.3560	0.0432	0.1479	0.2330	0.3349	0.4683	0.1096
ChatLaw	0.2049	0.2328	0.2745	0.0353	0.1306	0.1913	0.2684	0.3751	0.1285
SAILER	0.3142	0.4133	0.4745	0.0539	0.1780	0.3442	0.5688	0.7092	0.1427
GEAR	*	*	*	0.0630	0.1706	0.3142	0.4625	*	0.2162
BGE	0.4737	0.5539	0.5937	0.0793	0.2945	0.4298	0.6500	0.7394	0.1926
FET_{LLM}	0.5139	0.5862	0.6291	0.0980	0.2967	**0.4769**	0.6802	0.7828	0.2140
LLM	0.3583	0.4293	0.4798	0.0506	0.2240	0.3644	0.5383	0.6652	0.1453
$\text{FET}_{Notes\_guided}$	**0.5257**	**0.5987**	**0.6441**	**0.1073**	**0.3098**	0.4750	**0.6897**	**0.7974**	**0.2191**
Notes	0.3737	0.4602	0.5014	0.0730	0.2123	0.3586	0.5880	0.6798	0.1637

Table 5: SCR results. Bold fonts indicate leading results in each setting. * denotes that the indicator is not applicable to the current model. Since the output of GEAR cannot directly evaluate NDCG, the official results under the same setting are directly referenced in this paper. *LLM* and *Notes* represent the results of retrieval using only the four elements.

query and candidate, we prompt the LLM with the concept of the FET to generate case-specific four elements. During retrieval, the query and candidate are matched based on their case facts and the generated four elements, with scores from both components weighted accordingly. Based on testing, we assign a 7:3 ratio to the case facts and the four elements to balance detailed facts with theoretical key elements.

$\text{FET}_{\text{Notes\_guided}}$: This method leverages expert notes to guide the LLM in generating the four elements of a case, aiming to incorporate practical domain-knowledge. Unlike FET_{LLM}, this method first employs a smaller model to predict the case charges, retrieves notes associated with these charges, and uses them to guide the LLM in generating the four elements. This differs from prior methods of using notes in SCD, as the notes are used directly without refinement, posing greater challenges for the LLM.

6.2 Dataset

LeCaRDv2(Li et al., 2024c) is the latest version of LeCaRD(Ma et al., 2021), which is widely used in Legal Case Retrieval (LCR) (Li et al., 2024b; Zhou et al., 2023). It comprises 800 queries and 55,192 candidates extracted from 4.3 million criminal case documents. There are two common evaluation settings for this dataset: one uses a subset (Qin et al., 2024) with a candidate pool size of 1,390, while the other uses the full set (Li et al., 2024c) with a candidate pool size of 55,000. We conducted experiments under both settings.

Following previous work(Feng et al., 2024; Qin et al., 2024), we adopt commonly used evaluation metrics. For the subset, we use NDCG@10, 20, 30, Recall@1, 5, 10, 20, and MRR. For the full dataset, we use Recall@100, Recall@200, Recall@500, and Recall@1000.

6.3 Baselines

Consistent with earlier work(Li et al., 2024c; Qin et al., 2024), we compare two groups of baselines, Sparse retrieval methods and Dense retrieval methods, including: QL(Zhai et al., 2008), BM25(Robertson et al., 2009), BERT(Devlin, 2018), Lawformer(Xiao et al., 2021), ChatLaw-Text2Vec[1](Cui et al., 2023), SAILER(Li et al., 2023), GEAR(Qin et al., 2024).

Details of each baseline is shown in Appendix D. These baselines are implemented using the FlagEmbedding Toolkit[2]. All experiments were run on a server with a single RTX 3090.

6.4 Results

As shown in Table 5, the results analysis is as follows:

FET Works Well in LCR. The baseline model BGE achieves strong performance across most metrics compared to previous methods. Introducing the Four-Element Theory (FET) further improves its results, with relative MRR improvements of 11.11% for FET_{LLM} and 13.76% for $\text{FET}_{\text{Notes\_guided}}$, indicating that introducing legal theory is important to improve the performance of the model on LCR.

[1] https://modelscope.cn/models/fengshan/ChatLaw-Text2Vec

[2] https://github.com/FlagOpen/FlagEmbedding

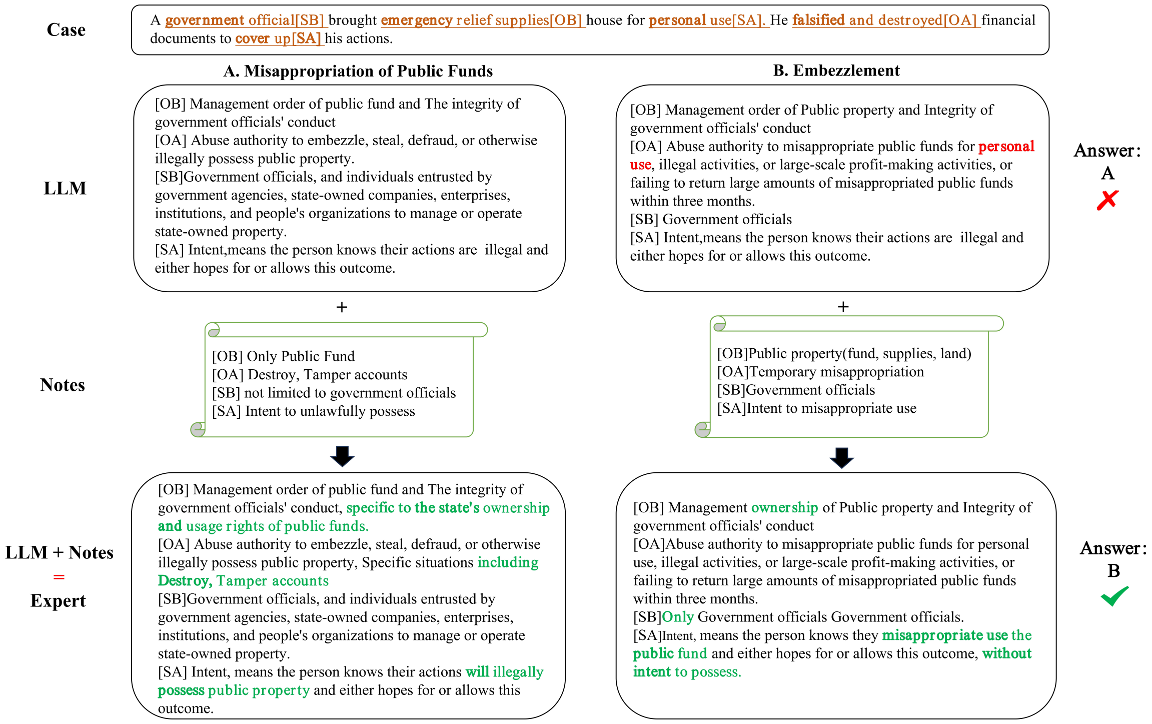

Case
A government official[SB] brought emergency relief supplies[OB] house for personal use[SA]. He falsified and destroyed[OA] financial documents to cover up[SA] his actions.
A. Misappropriation of Public Funds
B. Embezzlement
LLM
[OB] Management order of public fund and The integrity of government officials' conduct
[OA] Abuse authority to embezzle, steal, defraud, or otherwise illegally possess public property.
[SB]Government officials, and individuals entrusted by government agencies, state-owned companies, enterprises, institutions, and people's organizations to manage or operate state-owned property.
[SA] Intent,means the person knows their actions are illegal and either hopes for or allows this outcome.
[OB] Management order of Public property and Integrity of government officials' conduct
[OA] Abuse authority to misappropriate public funds for personal use, illegal activities, or large-scale profit-making activities, or failing to return large amounts of misappropriated public funds within three months.
[SB] Government officials
[SA] Intent,means the person knows their actions are illegal and either hopes for or allows this outcome.
Answer:
A
✗
+
Notes
[OB] Only Public Fund
[OA] Destroy, Tamper accounts
[SB] not limited to government officials
[SA] Intent to unlawfully possess
[OB]Public property(fund, supplies, land)
[OA]Temporary misappropriation
[SB]Government officials
[SA]Intent to misappropriate use
LLM + Notes
=
Expert
[OB] Management order of public fund and The integrity of government officials' conduct, specific to the state's ownership and usage rights of public funds.
[OA] Abuse authority to embezzle, steal, defraud, or otherwise illegally possess public property, Specific situations including Destroy, Tamper accounts
[SB]Government officials, and individuals entrusted by government agencies, state-owned companies, enterprises, institutions, and people's organizations to manage or operate state-owned property.
[SA] Intent, means the person knows their actions will illegally possess public property and either hopes for or allows this outcome.
[OB] Management ownership of Public property and Integrity of government officials' conduct
[OA]Abuse authority to misappropriate public funds for personal use, illegal activities, or large-scale profit-making activities, or failing to return large amounts of misappropriated public funds within three months.
[SB]Only Government officials Government officials.
[SA]Intent, means the person knows they misappropriate use the public fund and either hopes for or allows this outcome, without intent to possess.
Answer:
B
✓

Figure 3: A case of Embezzlement in SCD task. LLM-generated four elements led to the incorrect charge (Misappropriation of public funds) based on personal use. After adding notes, the model then identified the correct charge (Embezzlement) based on subtle differences (the green parts).

Notes Guidance is Necessary. By leveraging external annotations, FET$_{\text{Notes\_guided}}$ achieves significant improvements across most metrics, with an average gain of over 2.88%. Specifically, using notes-guided four elements (*Notes*) outperforms LLM-generated four elements (*LLM*) by an average of 12.66% in MRR, demonstrating the critical role of human expert knowledge in enhancing retrieval precision. A case study in Appendix F further supports this finding, showing that expert notes, though fragmented, provide practical judgment points and key narratives (e.g., establishing the Crime of Denuding Woods need to reach a big amount) that help the LLM focus on essential facts and refine case construction.

We also evaluated the FET method on the full dataset, as shown in Table 9. The results show that even when the candidate pool is expanded from 1.3k to 55k, the performance remains similar to previous results, with the notes-guided method still performing the best.

7 Discussion

How notes and LLM internel knowledge fuse? Figure 3 shows a case study. Although the elements generated by the LLM are standardized, they lack representativeness for the two charges, and the key points and distinctions are unclear. This led to the incorrect charge (Misappropriation of public funds) based on "personal use" in the task of SCD. After adding notes summarizing cases and knowledge from ordinary people, the notes-guided four elements improved in representativeness and standardization, making them comparable to the expert-generated four elements. The model then identified the correct charge (Embezzlement) based on subtle differences, such as whether accounts were destroyed.

8 Conclusion and Future Work

The expert notes proposed in this study show significant potential for application in other domains, such as medicine or finance, due to their ability to leverage expert knowledge at low cost. However, further research is needed to explore these possibilities. In fields involving sensitive data, such as medical records or prescriptions, careful consideration of ethical and privacy issues will be crucial.

9 Ethical Considerations

The datasets used in our evaluation are sourced from publicly available legal datasets, with all defendant information anonymized to ensure privacy.

Our work aims to explore how LLMs and human experts can better assist each other. Expert notes represent a scenario for efficiently utilizing informal expert knowledge. However, the expert annotations and notes in this paper are simulated through manual annotation and do not involve specific private information, such as identifiable individuals or events.

10 Limitations

While the annotations used in our experiments were created by annotators with a legal academic background, they are not practicing lawyers with extensive field experience. This gap occasionally led to some errors, such as misidentifying the subjects in the Four-Element Theory or confusing key details of low-frequency charges. These factors may introduce noise into the dataset and limit the framework's performance.

Acknowledgments

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2309.07597.

Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2454–2469.

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

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

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485.

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.

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

Ang Li, Qiangchao Chen, Yiquan Wu, Ming Cai, Xiang Zhou, Fei Wu, and Kun Kuang. 2024a. From graph to word bag: Introducing domain knowledge to confusing charge prediction. *arXiv preprint arXiv:2403.04369*.

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.

Haitao Li, Qingyao Ai, Xinyan Han, Jia Chen, Qian Dong, Yiqun Liu, Chong Chen, and Qi Tian. 2024b. Delta: Pre-train a discriminative encoder for legal case retrieval via structural word alignment. *arXiv preprint arXiv:2403.18435*.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024c. Lecardv2: A large-scale chinese legal case retrieval dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260.

Genlin Liang. 2017. The vicissitudes of chinese criminal law and theory: A study in history, culture and politics. *Peking University Law Journal*, 5(1):25–49.

Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. *arXiv preprint arXiv:2104.09420*.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.

Weicong Qin, Zelin Cao, Weijie Yu, Zihua Si, Sirui Chen, and Jun Xu. 2024. Explicitly integrating judgment prediction with legal document retrieval: A law-guided generative approach. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2210–2220.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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.

Shunyu Yao, Qingqing Ke, Qiwei Wang, Kangtong Li, and Jie Hu. 2024. Lawyer gpt: A legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, pages 108–112.

Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Pengwei Yan, Changlong Sun, Xiaozhong Liu, et al. 2024. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. *arXiv preprint arXiv:2410.02507*.

ChengXiang Zhai et al. 2008. Statistical language models for information retrieval a critical review. *Foundations and Trends® in Information Retrieval*, 2(3):137–213.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

Youchao Zhou, Heyan Huang, and Zhijing Wu. 2023. Boosting legal case retrieval by query content selection with large language models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 176–184.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model. *arXiv preprint arXiv:2406.04614*.

A Detailed Information on Notes Annotations

Each annotator underwent two rounds of training, and was provided with annotation samples and annotation instructions regarding each element (as shown in Table 6). The annotations made by each annotator were checked and revised at least twice.

B Human Evaluation Guidance

The annotators included three postgraduate students specializing in criminal law (different from the annotators in Section **??**) and one master's student in legal science and technology. The annotators scored independently, without knowledge of each other's results. Before scoring, they were asked to read the descriptions and scoring guidelines (as shown in Table 7) for each evaluation dimension. In order to ensure the fairness of the evaluation, they do not know the source of each four elements, and even do not know that these four elements include those generated by LLMs.

When assigning scores, they were also required to provide brief justifications. For example, for the Completeness dimension: 3 (The description of Objective Aspect is too brief, and does not specify the intent of illegal possession).

C Baselines in Similar Charge Disambiguation

Traditional Methods: To compare the performance of traditional methods and LLMs on SCD tasks, we evaluate several baseline models commonly used in previous work (Liu et al., 2021):

GCI (Liu et al., 2021) is a graph-based causal inference framework that constructs causal graphs from fact descriptions to assist legal decision-making.

CausalChain combines GCI with neural networks (NN) to capture crime patterns and represent the process of committing crimes.

Bi-LSTM (Zhou et al., 2016) serves as a representative backbone model for legal judgment prediction. We include three variants as baselines: standard Bi-LSTM, Bi-LSTM+Att (with attention mechanisms), and Bi-LSTM+Att+Cons, which incorporates legal constraint-based attention.

Element	Definition
Object	The embodiment of some abstract social interests. For example, the object of infringement of personal interests is the right to life, and the object of infringement of property interests is mobile phones, wallets, etc.
Objective aspect	The objective facts of criminal activity, including the key actions that triggered the crime, such as theft and robbery, and the results caused by the actions, such as serious injury, death, and property damage
Subject	The person who [commits a criminal act] and should [bear criminal responsibility] according to law. It is usually a general subject, but there are special subjects, such as the subject of the crime of corruption is the state staff.
Subjective aspect	The psychological attitude of the criminal subject towards the behavior endangering the society and the harm result caused by it. It is usually intentional or negligent.

Table 6: Definition of each element.

LLM-based Methods: For LLM baselines, we evaluate both general-purpose and task-specific methods.

GPT-4o is an optimized version of GPT-4(Achiam et al., 2023) that has well performance in specific tasks through domain adaptation.

To explore the effectiveness of notes-guided four elements in LLMs, we further consider other methods that introduced the Four-element theory into LLMs.

GPT-4o$_{\mathbf{Law}}$, which introduces articles related to corresponding charges into the instruction to provide legal context.

Legal-COT is a variant of COT (Kojima et al., 2022) that guides the LLM to perform step-by-step legal reasoning by incorporating explanations of the Four-element theory into the instruction.

As shown in Table 8, different methods differ in their prompts for generating and explaining the Four-Element Theory, but generally follow a similar process. For the SCD output, except for COT, which requires a step-by-step reasoning process and prediction results, all other methods only require the output of prediction results. In all the experiments on LLMs mentioned in this paper, the max_tokens of output is 3,000, and the temperature is set to 0 or 0.0001 (in multiple repeated experiments).

D Baselines in Legal Case Retrieval

Sparse Retrieval Methods: **QL**(Zhai et al., 2008) is a probabilistic retrieval model that ranks documents by the relevance likelihood to the query. **BM25**(Robertson et al., 2009) is a probabilistic retrieval model that calculates the doc-query relevance using term frequency and document length.

Dense Retrieval Methods: **BERT**(Devlin, 2018) is a language model widely used in retrieval tasks. In this paper, we chose BERT-base-Chinese[3]. **Legal-BERT**[4](Chalkidis et al., 2020) is a variant of BERT that is specifically trained on legal corpora. **Lawformer**(Xiao et al., 2021)is a Chinese legal pre-trained model based on Longformer(Beltagy et al., 2020), which is able to process long texts in the legal domain. **ChatLaw-Text2Vec**[5](Cui et al., 2023) is a Chinese legal LLM trained on 936,727 legal cases for similarity calculation of legal-related texts. **SAILER**(Li et al., 2023) is a structure-aware legal case retrieval model utilizing the structural information in legal case documents. **GEAR**(Qin et al., 2024) is a generative retrieval framework that explicitly integrates judgment prediction with legal document retrieval in a sequence-to-sequence manner.

E SCR results on the full LeCaRDv2 Dataset

As presented in Table 9, we selected several representative methods based on sparse retrieval and dense retrieval for experiments on the full LeCaRDv2 dataset. All language models were not fine-tuned. The notes-guided FET method achieved the best performance among all language models, attaining top results in both R@500 and R@1000. The results indicate that the conclusions drawn from the full dataset are consistent with those from the subset, and the notes-guided method demonstrates strong performance.

[3] https://huggingface.co/google-bert/bert-base-chinese

[4] https://github.com/thunlp/OpenCLaP

[5] https://modelscope.cn/models/fengshan/ChatLaw-Text2Vec

Dimension	**Precision**	**Completeness**	**Representativeness**	**Standardization**
Definition	Whether there are errors in key elements	Whether the four elements are complete	Whether key elements and scenarios are emphasized	Whether language and format are clear and standardized
Score 1	Contains numerous obvious errors, severely impeding the judgment of culpability, exculpation, and conviction, leading to significant deviations.	Severe omission of key content, unable to present a complete picture of the crime structure, greatly hindering analysis of criminal behavior.	Completely fails to mention any key elements or scenarios, unable to highlight essential points for crime recognition, offering no assistance in conviction.	Language is extremely chaotic and obscure; format lacks any standardization, greatly hindering comprehension and application.
Score 2	Contains multiple noticeable errors, significantly interfering with culpability, exculpation, and conviction judgments, potentially leading to partial errors.	Noticeable omissions in content, failing to comprehensively cover crime elements, affecting thorough analysis of criminal behavior.	Only highlights a minimal and unimportant portion of the key elements, providing weak support for understanding key crime features.	Language is relatively vague and inaccurate, with a casual format that makes content comprehension significantly challenging.
Score 3	Contains a few errors, but the overall accuracy in determining culpability, exculpation, and conviction is relatively unaffected, unlikely to lead to judgment errors.	Some key content descriptions are incomplete, but they generally present the framework of the crime structure.	Highlights some relatively important key elements but lacks comprehensiveness and prominence, offering limited assistance in crime identification.	Language is generally clear but may have minor deviations in phrasing or formatting.
Score 4	Almost error-free, key elements accurately serve culpability, exculpation, and conviction judgments, ensuring the accuracy of results.	Key elements are mostly complete, with only very slight and non-critical deficiencies that do not hinder a comprehensive analysis of the crime.	Clearly and relatively comprehensively highlights key elements, aiding in accurately identifying crucial aspects of criminal behavior.	Language is clear and accurate, format is relatively standardized, facilitating comprehension and application of relevant content.
Score 5	Completely error-free, key elements are precisely defined, achieving highly accurate culpability, exculpation, and conviction judgments without any flaws.	All four elements are complete and detailed, covering every aspect of the crime, perfectly presenting the crime structure.	Precisely and comprehensively highlights all crucial elements, enabling immediate grasp of the core aspects of the crime, significantly aiding conviction.	Language is extremely clear, standardized, and concise; format perfectly meets requirements, with no barriers to understanding, ensuring efficient information delivery.

Table 7: The four dimensions of the human evaluation and the specific score description.

F A Case Study of LCR

Table 10 presents a case study involving the Crime of Denuding Woods. By comparing the original expert notes, the LLM-generated four elements of the case, and the Notes-guided LLM-generated results, we observe the following:

1) The large language model demonstrates an inherent ability to identify important aspects based on its internal knowledge. For instance, in the Objective Aspect, the LLM highlights “cut down trees without permission.” After integrating the expert notes, this detail is retained, reflecting the model’s independent judgment.

2) Incorporating expert notes enables the model to better emphasize conviction- and sentencing-related factors (e.g., establishing the Crime of Denuding Woods need to reach a big amount). It also enhances the precision of key case descriptions, such as specifying “the total of 4 times”, which is crucial for matching cases with similar facts.

Method	**GPT-4o**	**GPT-4o$_{\textbf{Law}}$**	**Legal-COT**	**FET$_{\textbf{LLM}}$**	**FET$_{\textbf{Expert/Notes}}$**	**FET$_{\textbf{Notes\_guided}}$**
Pre-task	None	None	None	LLM-generated four elements	Expert-annotated or Notes-type four elements	Notes-type four elements
Prompt	You are a lawyer specializing in criminal law. Based on Chinese criminal law, please determine which of the following candidate charges the given facts align with.					
	Candidate charges are as follows: *#Candidate Charges*	The candidate charges and relevant legal articles are as follows: *#Candidate Charges + #Articles*	Please analyze using the Four Elements Theory step by step: *#details about each step.* The candidate charges are as follows: *#Candidate Charges*	The candidate charges and their corresponding four elements are as follows: *#Four Elements of candidate charges*. The four elements represent the four core factors of a charge. Compare the case facts to determine which charge's four elements they align with, thereby identifying the charge.		
	Output format: *#Format*. Note: Only output the charge, no additional information. Case facts: *#Case Facts.*					

Table 8: Prompts of different methods in Similar Charge Disambiguation. # represents a format input.

Model	**R@100**	**R@200**	**R@500**	**R@1000**
BM25	**0.6262**	**0.6629**	0.6949	0.7207
QLD	0.5984	0.6576	0.7065	0.7424
BERT	0.1116	0.1493	0.2174	0.2819
Lawformer	0.2432	0.304	0.4054	0.4833
ChatLaw	0.1045	0.1628	0.2791	0.3999
SAILER	0.2834	0.4033	0.6104	0.7568
BGE	0.4085	0.5246	0.6855	0.7912
FET$_{\text{LLM}}$	0.4167	0.5388	0.7006	0.7925
FET$_{\text{Notes\_guided}}$	0.4199	0.5393	**0.7012**	**0.7930**

Table 9: SCR results on the full set of LeCaRDv2. Bold fonts indicate leading results in each setting. All language models are not fine-tuned, so BM25 and QLD are strong baselines. The notes-guided FET method achieved the best performance among all language models and attained the top results in both R@500 and R@1000.

Document	*[Head of document]*...The People's Procuratorate of XX County alleged that between 201X and 201X, the defendants A and B engaged in raising cattle in XX County. During this period, to fence off free-range cattle in the forest, facilitate passage, and repair cattle sheds, A and B, without obtaining a logging permit, independently or jointly cut down trees on collective public hills and villagers' private forests in XX Village, XX Town, XX County. The illegal logging occurred on four occasions, totaling 633 trees, with a cumulative timber volume of 42.306 cubic meters. Here are the facts...		
Charge	Crime of Denuding Woods		
	Notes	**LLM-generated**	**Notes-guided**
Four Elements	**Object:** The state's regulatory activities over forest resources, including trees on construction land, trees on national highways, and farmland shelterbelts. **Objective Aspect:** Deforestation, chopping, big amount **Subject:** General subject **Subjective Aspect:** Deliberate intention	**Object:** Environmental resources interests: forest resources **Objective Aspect:** The defendants A and B have repeatedly cut down trees without permission, cutting down a total of 633 trees, with a total standing stock of 42.306 cubic meters. **Subject:** Defendants A and B, both natural persons with full criminal responsibility. **Subjective Aspect:** Deliberate intention	**Object:** The state's regulatory activities over forest resources **Objective Aspect:** A and B, in the absence of forest harvesting permits, felled trees a total of 4 times, in big quantities, amounting to 633 trees with a total standing volume of 42.306 cubic meters. **Subject:** Defendants A and B, both natural persons with full criminal responsibility. **Subjective Aspect:** Deliberate intent, demonstrated by awareness of the lack of a logging permit and willful engagement in unauthorized logging.

Table 10: The results of the case four elements obtained through different methods in LCR. **Notes** refer to expert notes related to the charge retrieved during the search. **LLM-generated** and **Notes-guided** indicate whether using expert notes to guide LLM in generating the four elements. Red parts mean the knowledge from the expert notes, while blue parts show the LLM's internal knowledge. By incorporating the expert notes, the model better emphasizes conviction and sentencing related information (e.g., big amount) and provides more detailed descriptions of critical case facts (e.g.,4 times).