JUREX-4E: Juridical Expert-Annotated Four-Element Knowledge Base for Legal Reasoning

Anonymous ACL submission

Abstract

The Four-Element Theory is a fundamental framework in criminal law, defining the constitution of crime through four dimensions: Subject, Object, Subjective aspect, and Objective aspect. This theory is widely referenced in legal reasoning, and many Large Language Models (LLMs) attempt to incorporate it when handling legal tasks. However, current approaches rely on LLMs' internal knowledge to incorporate this theory, often lacking completeness and representativeness. To address this limitation, we introduce JUREX-4E, an expert-annotated knowledge base covering 155 criminal charges. It is structured through a progressive hierarchical annotation framework that prioritizes legal source validity and employs diverse legal interpretation methods to ensure comprehensiveness and authority. We evaluate JUREX-4E on the Similar Charge Distinction task and apply it to Legal Case Retrieval, demonstrating its effectiveness in improving LLM performance. Experimental results validate the high quality of JUREX-4E and its substantial impact on downstream legal tasks, underscoring its potential for advancing legal AI applications.

1 Introduction

In legal AI tasks, enhancing the accuracy and interpretability of Large Language Models (LLMs) in the legal domain often requires the incorporation of legal theories as a support (Jiang and Yang, 2023; Servantez et al., 2024; Yuan et al., 2024; Deng et al., 2023). One important theory is the Four-Element Theory of Crime Constitution in Chinese criminal law (Liang, 2017). This theory deconstructs criminal conduct into four elements: Subject, Object, Subjective aspect, and Objective aspect, providing clear standards for judicial authorities to determine criminal behavior and helping to prevent the abuse of penal power.

However, most current approaches do not provide additional knowledge but rather rely on the LLM's internal knowledge to incorporate the Four-Element Theory. A common method is to guide LLMs in mimicking expert reasoning processes. For example, designing four separate prompts to guide the LLM outputs in the form of four elements(Deng et al., 2023).

These methods assume that the model has a solid grasp of the Four-Element Theory, which has not yet been verified. We had LLMs generate the four elements of several complicated crimes in Chinese judicial practice(Ouyang et al., 1999), and then asked legal experts to score them. We found that, although LLMs can generate formally standardized and relatively accurate legal descriptions when provided with legal theoretical frameworks and references, the model still underperformed in terms of completeness and representativeness. This shortcoming could affect the accuracy and soundness of subsequent reasoning.

To help LLMs better utilize the Four-Element Theory in legal tasks, we propose **JUREX-4E: JUR**idical **EX**pert-annotated **4-E**lement knowledge base for legal reasoning. This knowledge base is annotated using a progressive hierarchy: Article → Judicial Interpretations → Guiding Cases → Academic Discourses, which is built upon the pyramid structure of legal source validity. It incorporates various legal interpretation methods, including textual, systematic, sociological, and purposive interpretations. The knowledge base covers the four elements of 155 high frequency charges, annotated by legal experts over a period of seven months. Each crime's four elements are described in an average of 472.5 words.

To assess the quality of the annotations, we sampled several crimes for human evaluation. The expert annotations achieved an average score of 4.60 on a 5-point scale, while the LLM-generated four elements scored only 3.96, indicating that the expert annotations were of higher quality. To further evaluate the annotations objectively and compre-

hensively, a direct way is to judge whether different charges can be distinguished according to the four-element definition of crime constitution. Therefore, we introduced the Similar Charge Distinction task (Liu et al., 2021). For each case, we provided the four elements of the candidate confused charges and combined them with the case facts as model input. The experimental results showed that injecting expert annotations helped the model better differentiate between similar charges, improving performance with a 0.65 increase in average accuracy and a 0.70 increase in average F1-score, underscoring the superior quality and reliability of expert annotations compared to those generated by the LLM.

We also applied the expert annotations in a specific legal task: Legal Case Retrieval. It is an important step in the practice of analyzing cases and making judgments, requiring the precise application of the four-element theory to compare the criminal composition of cases. We designed a simple retrieval framework guided by expert knowledge, in which the charge's four elements was used to generate four-element descriptions for both the query and candidate cases, and then match similar cases based on their vector similarity. Experiments demonstrated that incorporating expert-annotated four elements improved retrieval performance, as the model became better at focusing on the legal features and key details.

Our contributions are as follows:

(1) We verify that LLMs have gaps in understanding the legal theory, highlighting the inadequacy of relying solely on LLM-driven reasoning for legal AI tasks.
(2) We built the JUREX-4E knowledge base, which is the first to incorporate the pyramid structure of legal source validity and covers the four elements of 155 criminal charges under Chinese Criminal Law.
(3) We demonstrated the significance of incorporating criminal composition elements in the Similar Charge Distinction task and proved the superior quality of the expert-annotated four-element knowledge base.
(4) We applied JUREX-4E to the Legal Case Retrieval task, found that they do indeed contribute to downstream tasks.

2 Related Work

In legal AI, much work has introduced legal theories to enhance reasoning and improve model accuracy and interpretability. For example, legal syllogism prompting (LoT)(Jiang and Yang, 2023) teaches LLMs for legal judgment prediction by instructing legal syllogism, Chain of Logic(Servantez et al., 2024) guides models in reasoning about compositional rules by decomposing logical statements based on the IRAC (Issue, Rule, Application, Conclusion) paradigm. Among these, the Four-Elements Theory (FET) of Crime Constitution is a widely adopted framework(Yuan et al., 2024; Deng et al., 2023).

The Four-Element Theory is one of the most widely recognized criminal theories 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 can be briefly summarized 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.

Before discussing the Four-Element Theory (FET), it is necessary to briefly compare it with another key theory in Chinese criminal law, the Hierarchical Theory of Crime Constitution(Zhou, 2017b; Zhang, 2010), the main distinction between these theories lies in whether a hierarchical structure is considered, with ongoing debates in practice(Gao, 2009; Chen, 2010, 2017; Zhou, 2017a). We chose FET as our foundational template for following reasons: 1) its dominance in Chinese judicial practice aligns with real-world criminal judgments; (2) its clear distinction between objective aspects and subjective intent offers direct reasoning checkpoints compared to the Three-Tier Theory; (3) its four-element annotation is flexible and can be adapted to the Three-Tier Theory by prioritizing objective analysis before subjective evaluation(Li, 2006; Zhang, 2017).

Recent approaches have leveraged the FET framework to model expert reasoning. For exam-

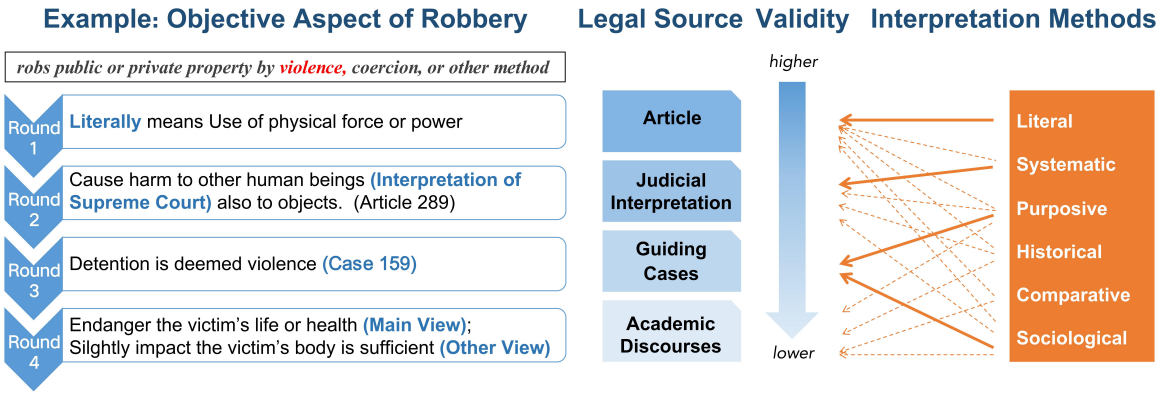

Figure 1: Hierarchical Legal Interpretation System base on legal source validity. The legal sources follow a hierarchical order of validity. Thick arrows indicate the primary level where a particular interpretive method is applied, while dashed arrows represent its supplementary use at that level.

ple, breaking down legal rules into FET-aligned components using automated planning techniques (Yuan et al., 2024). Employing model-generated four-element structures as minor premises in legal judgment analysis (Deng et al., 2023). While these methods have demonstrated improved performance on downstream tasks, they generally assume that the LLMs inherently understand the FET, without systematically validating this assumption. Notably, prior research on criminal charge prediction (An et al., 2022) suggests that the models may misinterpret key legal concepts and may not be sensitive enough to the subtle differences in fact descriptions of confusing charges, highlighting the need to incorporate expert annotations to support LLM reasoning.

3 Dataset Construction

3.1 Hierarchical Legal Interpretation System

Annotating the four elements of crime constitution is essentially a process of legal interpretation, which can be analyzed in two aspects:

(1) **What law is being interpreted.** This involves identifying the sources of law, including statutory provisions corresponding to a specific charge, their associated judicial interpretations, case precedents, and academic discourses. In legal studies, these sources are categorized based on their legal validity into formal sources (which carry legal forces in judgments) and informal sources (which serve as references without legal forces)(Pound, 1925; Watson, 1982; Pound, 1932). Articles and judicial interpretations are considered formal sources, whereas case precedents and academic discourses are regarded as informal sources under the Chinese legal system(Zhang and Zhou, 2007).

(2) **How the law is interpreted.** This pertains to legal interpretation methods, including literal interpretation, systematic interpretation, purposive interpretation, etc. These methods follow a hierarchical order in legal reasoning(Sutherland, 1891; Kim and Division, 2008; Eig Larry, 2014). Legal interpretation should begin with literal interpretation (textual analysis). If the intended meaning cannot be clearly derived from the article alone, systematic interpretation and purpose interpretation should be applied first. If ambiguity remains, historical interpretation and comparative law interpretation may be used to further clarify the legal meaning. The specific definition is in AppendixB.

Based on these principles, our annotation follows a pyramid structure of Hierarchical Legal Interpretation System base on legal source validity. As shown in Figure 1, the system is divided into two parts: Legal Source and Legal Interpretation Methods. The main structure of legal source follows a hierarchical order of validity: Article → Judicial Interpretations → Guiding Cases → Academic Discourses, where various legal interpretation methods are applied across different levels. Thick arrows indicate the primary level at which a particular method is used, while thin arrows denote the cross applications.

3.2 Hierarchical Annotation Path of Legal Sources

Our Annotators are experts have all passed the National Uniform Legal Profession Qualification Examination and are familiar with the Four-Element Theory. The entire annotation process took a total of 7 months and involved 4 rounds of annotation according to the validity of the legal source from

high to low level.

The First Level: Article. Legal elements can be seen as an interpretation and refinement of the statutory provisions corresponding to a particular crime. Using **literal interpretation** as the primary method, the statute is broken down based on its semantic meaning and common usage, ensuring that the interpretation does not extend beyond the possible meaning of the text: (1) linguistic analysis follows the subject-predicate-object structure of the provision. (2) To maintain consistency, terms are systematically classified and mapped(e.g: subjective aspect is classified as either intentional or negligent.)(3) Only when it is impossible to make an explicit inclusion or exclusion judgment for an element based on the rules of language use (neutral option field), other interpretation methods should be used. This initial phase takes almost 2 months.

For example, in the crime of robbery, the object "public or private property" represents the protected legal interest. The phrase "forcibly seizing public or private property through violence, coercion, or other means" describes the objective aspect. Since no subject is specified, it is assumed to involve a general subject, and the adverbs "violence" and "coercion" indicate an intentional act. Preliminarily interpret 'violence' in the objective aspect as 'Use of physical force or power', but the specific forms and subjects of violence need further clarification.

The Second Level: Judicial Interpretation. In the 3rd and 4th months, the second level of the hierarchical annotation path focuses on refining legal elements through judicial interpretation. The primary method used for interpreting these materials is **systematic interpretation**. This approach examines the position of the corresponding articles within the legal system by analyzing their placement within the structure of laws, including parts, chapters, sections, articles, clauses, and subclauses, as well as their relationship to other statutes and judicial interpretations. Additionally, other interpretative methods, such as sociological interpretation and teleological interpretation, are referenced based on judicial interpretations, related statutory provisions, or bar exam questions.The goal of this level is to clarify the legislative intent by considering the contextual relevance of each provision within the broader legal framework.

For example, in the first level, the objective aspect of "violence" in the crime of robbery requires further clarification, specifically regarding whether violence must be directed exclusively at persons or could also apply to property. Article 289 of Chinese Criminal Law(Congress, 2017) stipulates that in cases of "smashing, looting, and robbing" committed by a group, the ringleaders shall be convicted of robbery if they destroy or seize public or private property. This provision demonstrates that violence against property can also constitute robbery under Chinese law.

The Third Level: Guiding Cases. In the 5th to 6th month, **purposive interpretation and sociological interpretation** are applied to the guiding cases and landmark judgments from the Supreme Court. By examining the social significance of real-world cases, these methods bridge the subtle gap between abstract legal theory and practical cases. This approach enables dynamic adaptation and integration of empirical insights and emerging controversies within the dataset.

For example, in Criminal Trial Reference Case No.159(Zou, 2002), the perpetrator lured the victim into a room, locked the door, and seized 170,000 RMB intended for a transaction. The court determined that although the detention did not endanger personal safety, it was sufficient to suppress the victim's resistance, thus constituting "violence" in in the objective aspects of robbery. Another example is the "Molestation and Theft Case" (Ma, 2021), where the perpetrator bound the victim, committed molestation, and stole the victim's phone. Since the ongoing molestation reinforced coercion, it constitutes a new act of violence. Thus, the annotation includes "molestation" as an additional method.

The Fourth Level: Academic Discourses. In the 7th month, the final stage involves academic expansion. Academic controversies are introduced by employing multiple interpretive methods such as **comparative law interpretation, purposive interpretation, and sociological interpretation**. These methods include inserting conflict markers at key points of controversy, highlighting the distinctions between mainstream consensus and minority theories, while providing brief annotations of their legal reasoning. This approach ensures the extensibility and academic depth of the dataset.

For example, regarding the crime of robbery, for the main view in China, Soviet Union, North Korea, and Japan explicitly holds that the violence must be severe enough to endanger the victim's life or health(Zhang, 2007). But some scholars

argue that any violence that can forcibly impact the victim's body is sufficient to constitute violence in robbery, no need to endanger the victim's life or health(Yang, 2010).

4 Data Distribution

Metric	LLM		Expert	
	Mean	Median	Mean	Median
Avg. Length	115.43	-	472.53	-
SB	23.12	27	51.64	17
OB	15.86	15	36.01	25
SA	28.00	30	42.38	21
OA	48.45	45	342.5	230

Table 1: Comparison of Legal Element Lengths: LLM vs. Expert. SB = Subject, OB = Object, SA = Subjective Aspect, OA = Objective Aspect.

As shown in Table 1, we compare the length of legal elements between expert-annotated descriptions in JUREX-4E and LLM-generated outputs across 105 charges that overlap with the Lecard-V2 dataset (Li et al., 2024c), which is one of the most comprehensive legal datasets, covering 184 criminal charges. We find that:

(1) The average total length of expert annotations (472.53) is more than four times longer than that of LLM-generated outputs (115.43), indicating that the former include more detailed information.

(2) The median difference between the Subject (SB), Object (OB), and Subjective Aspect (SA) is relatively small, as these elements are typically fixed. For example, the SB is often a general entity, and the SA is often intent or negligence.

(3) The median and mean values for SB and SA in the expert annotations differ, especially for SB (17 v.s. 51.64). This discrepancy arises because certain specialized charges may require more detailed explanations. For example, in the crime of copyright infringement, the definition of "work" under the subject element has 9 occasions. Detailed data distribution for each element is provided in Appendix A.

(4) The main difference between Expert and LLM is in the Objective Aspect (342.5 v.s. 48.45 in Mean). This is because the OA includes a range of factual elements describing the criminal behavior, such as the conduct, object, result, time, and location, which are most emphasized in legal provisions and are central to various legal interpretive theories.

5 Human Evaluation

We selected 6 complicated crimes in Chinese judicial practice(Ouyang et al., 1999) to evaluate whether the LLM can handle the Four-element Theory. Drawing from previous work(Deng et al., 2023; Cui et al., 2024; Zhou et al., 2023), we define LLM-generated knowledge as information produced by the LLM based on its pre-trained knowledge and contextual prompts. For detail, we provide the LLM with legal articles and the definition of each element in FET, prompting it to generate the four-elements base on these metrical. The LLM is expected to autonomously identify and generate the four elements based on its learned understanding of legal concepts.

We invite legal experts to assess the four elements generated by the LLM from four dimensions: **Precision, Completeness, Representativeness, and Standardization**:

- Precision: Whether the key components of each element are accurately identified. This dimension mainly evaluates whether the four elements faithfully represent the legal provisions.
- Completeness: Whether all necessary information of each element is included. This assesses whether any essential content is missing, such as the omission of a description for specific subjects, like government officials.
- Representativeness: Whether the annotations highlight the most critical scenarios in judicial practice. For example, in crimes of intentional injury, this would involve describing the representative means of harm.
- Standardization: Whether the four elements are clearly defined, ensuring consistency in the expression of identical elements across different crimes (e.g., consistent description of general subjects), with concise and easily understandable explanations, free from legal ambiguities or misunderstandings.

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 C.

As shown in Table 2, expert annotations consis-

Dimension	LLM	Expert	δ
Precision	4.12	4.69	+ 0.57
Completeness	3.79	4.65	+ 0.86
Representativeness	3.60	4.48	+ 0.88
Standardization	4.33	4.56	+ 0.23

Table 2: Performance comparison of four elements across methods. δ represents the score difference between expert and LLM-generated four-elements, with experts outperforming LLMs in all dimensions.

tently outperform LLM-generated elements across all four dimensions, highlighting the limitations of LLMs in understanding legal elements. The most pronounced deficiencies are observed in Completeness (+0.86) and Representativeness (+0.88). This suggests that while LLMs can generate formally standardized and relatively accurate four elements, their description are not specific enough and do not adequately reflect the representative features of a charge's criminal composition.

6 Evaluate Expert Knowledge on Charge Disambiguation

In the preceding section, the human evaluation demonstrated that experts annotated higher-quality four-elements. To further quantitatively evaluate the annotations, a direct way is to judge whether different charges can be distinguished according to the four-element definition of crime constitution. Therefore, we introduce the Similar Charge Disambiguation (SCD) task(Yuan et al., 2024; Li et al., 2024a).

6.1 Experiment Settings

6.1.1 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). Each crime has over 1.9k cases, with a total of 13,962 cases. The details of the classification groups are shown in Appendix D. Following previous work (Liu et al., 2021; Yuan et al., 2024), we use Average Accuracy (Acc) and macro-F1 (F1) as evaluation metrics.

6.1.2 Baselines and Methods

To evaluate SCD tasks, we consider two ways of incorporating legal knowledge. The first directly integrates legal statutes, represented by GPT-4o (Achiam et al., 2023) as the baseline and GPT-4o+Article, which explicitly provides relevant legal articles to the model. The second adopts structured legal reasoning to enhance interpretability and accuracy. We consider Legal-CoT, a Chain-of-Thought (Kojima et al., 2022) variant that conducts a stepwise analysis based on the FET, and MALR (Yuan et al., 2024), a multi-agent framework that decomposes legal tasks into sub-tasks in four-element structures. Details of each baseline are provided in Appendix D.

We use an unified approach to introduce four-element descriptions. For each group of similar charges, the model receives charges' four-elements from JUREX-4E or generated by LLM to aid classification. Specifically, GPT-4o+FET$_{\text{Expert}}$ relies on expert-annotated four-elements, while GPT-4o+FET$_{\text{LLM}}$ relies on LLM-generated four-elements. As shown in Appendix D, the instruction format is consistent across methods, with only the *[Four Elements of candidate charges]* varying based on the source. All experiments are conducted in a zero-shot setting, with the max_tokens set to 3,000 (or 10,000 for COT and MALR reasoning) and temperature set to 0 or 0.0001(In repeated experiments).

6.2 Results

As shown in Table 3, the GPT-4o+FET$_{\text{Expert}}$ performs best in discriminating similar charges, indicating that expert annotation is superior to other methods of directly or indirectly introducing FET with LLMs. Specifically, we can derive the following observation:

Effectiveness of Domain-Specific Legal Knowledge: Among all approaches, those that explicitly incorporate domain-specific legal knowledge, such as GPT-4o+Article, Legal-CoT, and MALR, outperform GPT-4o alone. This highlights the importance of integrating legal knowledge.

Importance of Concrete Four-element Knowledge: The accuracy of both Legal-CoT and MALR is still lower than GPT-4o+FET methods. This suggests that, compared to embedding the Four-Element Theory into LLMs' reasoning process, providing concrete charge four-elements en-

Model	F&E		E&MPF		AP&DD		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
GPT-4o	94.36	95.81	86.49	89.76	85.54	87.12	88.72	90.07
GPT-4o+Article	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
MALR	94.62	95.82	86.99	86.98	87.86	88.68	89.82	90.49
GPT-4o+FET$_{LLM}$	95.73	96.56	91.87	92.01	89.61	89.69	92.40	92.75
GPT-4o+FET$_{Expert}$	**96.06**	**96.69**	92.57	**93.05**	**90.53**	**90.62**	**93.05**	**93.45**

Table 3: Results of Charge Disambiguation. FET means introducing the Four-element theory with knowledge obtained from experts and LLM method. Highest results are in bold.

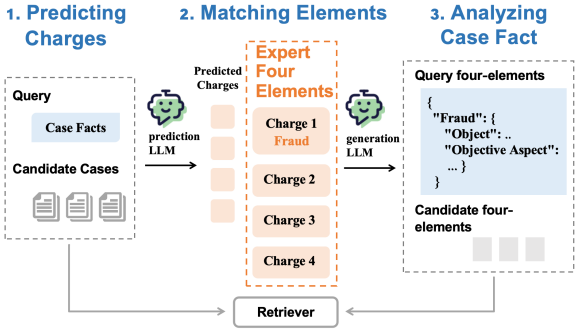

Figure 2: An expert-guided FET method to enhance legal case retrieval by incorporating expert four elements.

ables the model to better understand the different crimes' composition.

Superiority of Expert Annotations: Compared with the indirect introduction of FET reasoning, the method of directly introducing four-elements to the model (GPT-4o+FET) achieves better results. Notably, GPT-4o+FET$_{\text{Expert}}$ surpassing the GPT-4o+FET$_{\text{LLM}}$ by 0.65 in average accuracy and 0.70 in average F1-score, underscoring the superior quality and reliability of expert annotations in legal tasks, aligning with human evaluations in Table 2 and reaffirming the critical role of human expertise in legal decision-making.

7 Can Expert Knowledge Benefit More Downstream Tasks?

In this section, we design a simple framework to apply the expert-annotated four elements to Legal Case Retrieval (LCR), a task in which relevant cases are retrieved based on given 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 matches cases with similar criminal compositions.

7.1 Method

We implement a standard dense retrieval approach **BGE** using BGE-m3 (Chen et al., 2023), an advanced embedding model for dense retrieval. Given a query q and a candidate case c, , their vector representations $\mathbf{v}_q$ and $\mathbf{v}_c$ are obtained through shared encoder E: $\mathbf{v}_q = E(q), \quad \mathbf{v}_c = E(c)$. We used the BGE-m3 model without fine-tuning as the shared encoder. Next, the relevance score is computed via cosine similarity:

$$\text{sim}_{\text{base}}(q, c) = \frac{\mathbf{v}_q \cdot \mathbf{v}_c}{\|\mathbf{v}_q\|\|\mathbf{v}_c\|} \quad (1)$$

To retrieve the top-k most similar cases, we rank the candidates based on their cosine similarity to the query. Denote the set of candidate cases as $C = \{c_1, c_2, \ldots, c_n\}$, where n is the total number of candidate cases. We compute the similarity for each $c_i \in C$, and select the top-k candidates with the highest similarity scores.

As shown in Figure 2, to leverages expert-annotated four elements of charges, we introduce an **BGE+FET$_{\text{Expert_guided}}$** method for the retrieval process, consisting of three steps: (1) Predicting charges, a LLM $\mathcal{M}_p$ predicts potential charges $Z = \{z_1, ..., z_k\}$ from case facts. (2) Matching elements, retrieving corresponding charge's four-elements $\{f_z\}_{z \in Z}$ in JUREX-4E. (3) Analyzing case facts. Guided by $\{f_z\}$, another LLM $\mathcal{M}_g$ generates case-specific four elements a_c for candidate c. The final similarity score combines factual and theoretical alignment:

$$\text{sim}_{\text{final}}(q, c) = \alpha \cdot \text{sim}_{\text{base}}(q, c) + (1-\alpha) \cdot \text{sim}_{\text{f}}(a_q, a_c) \quad (2)$$

where $\alpha = 0.7$ and sim_{f} measures the similarity between the generated four-element descriptions.

To facilitate comparison, we also design a **BGE+FET$_{\text{LLM}}$** method that directly prompt the LLM $\mathcal{M}_g$ with the concept of Four-Element Theory to generate case-specific four elements a_c.

Model	NDCG@10	NDCG@20	NDCG@30	R@1	R@5	R@10	R@20	R@30	MRR
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
- base	0.3583	0.4293	0.4798	0.0506	0.2240	0.3644	0.5383	0.6652	0.1453
$\text{FET}_{Expert\_guided}$	**0.5211**	**0.5920**	**0.6379**	**0.1024**	**0.3049**	**0.4883**	**0.6885**	**0.7967**	**0.2155**
- base	0.3766	0.4584	0.5111	0.0715	0.1894	0.3709	0.5891	0.7203	0.1624

Table 4: SCR results. Bold fonts indicate leading results in each setting. * denotes that the indicator is not applicable to the current model.

7.2 Dataset

LeCaRDv2(Li et al., 2024c) is the latest version of LeCaRD(Ma et al., 2021), which is widely used in LCR task (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.

7.3 Baselines

Consistent with earlier work(Li et al., 2024c; Qin et al., 2024), we compare some dense retrieval methods, including: 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 E. These baselines are implemented using the FlagEmbedding Toolkit[2] with a RTX 3090.

7.4 Results

The LCR results are shown in Table 4, where we can observe that:

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 11.89% for $\text{FET}_{\text{Expert\_guided}}$, indicating that introducing legal theory is effective.

Expert Knowledge is Necessary. By leveraging external knowledge, $\text{FET}_{\text{Expert\_guided}}$ achieves significant improvements across all of the metrics. Specifically, using expert-guided case four-elements ($\text{FET}_{\text{Expert\_guided}}$-*base*) outperforms LLM-generated case four-elements (FET_{LLM}-*base*) by an average of 11.77% in MRR, demonstrating the critical role of expert knowledge in enhancing retrieval precision. A case study in Appendix G shows that the expert four-element for charges provide practical judgment points and key narratives (e.g., the special subject of the Crime of Embezzlement) that help the LLM focus on essential facts to analyze the case.

We also evaluated the FET method on the full set, as shown in Table 9 , and the results remain consistent, with the expert-guided method still performing best.

8 Conclusion

In this paper, we propose an expert-annotated knowledge base, evaluate its quality in the Similar Charge Distinction task, and apply it to the Legal Case Retrieval task. Our results demonstrate that expert annotations significantly enhance LLMs' understanding of the Four-Element Theory. The four-element annotations, enriched with professional legal interpretations, provide strong support for LLMs' reasoning capabilities. This approach can be extended to other legal AI tasks, such as legal document analysis and contract interpretation.

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

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

9 Ethical Considerations

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

10 Limitations

As a limitation, this knowledge base focuses on the Four-Element Theory within the context of 155 crimes under Chinese Criminal Law. However, the four-level hierarchical pyramid annotation structure based on the legal interpretation system proposed in this work provides valuable insights for future expansion to other legal domains, as it represents a theoretical framework in the field of jurisprudence. The interpretative methods within the legal interpretation system, including textual, systematic, sociological, and doctrinal interpretations, are universally recognized in international law field and can be applied to different laws, countries, and legal systems.

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*.

Zhenwei An, Quzhe Huang, Cong Jiang, Yansong Feng, and Dongyan Zhao. 2022. Do charge prediction models learn legal theory? *arXiv preprint arXiv:2210.17108*.

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.

Xingliang Chen. 2010. Crime constitution theory: An academic historical investigation from the four-elements to the three-tier theory. *Peking University Law Journal*, 22(1):49–69.

Xingliang Chen. 2017. Criminal law hierarchical theory: A comparative study between the three-tier and four-elements theories. *Tsinghua Law Journal*, 5.

National People's Congress. 2017. *Criminal Law of the People's Republic of China*. China Legal Publishing House.

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*.

M Eig Larry. 2014. Statutory interpretation: General principles and recent trends. *Congressional Center for Research*, (s 37).

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.

Mingxuan Gao. 2009. On the rationality of the four-elements theory of crime constitution and adherence to china's criminal law system. *China Legal Sci.*, (2):5–11.

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.

Yule Kim and American Law Division. 2008. *Statutory interpretation: General principles and recent trends*. Congressional Research Service Washington, DC.

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.

Hong Li. 2006. No need to reconstruct china's crime constitution system. *Chinese Journal of Law*, (1):32–51.

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*.

Yinxiang Ma. 2021. The spiritualization and limitation of the concept of violence in robbery. *Law Science*, (06):76–91.

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.

Tao Ouyang, Kejia Wei, and Renwen Liu. 1999. Confusing crimes, noncrime, and boundaries between crimes.

Roscoe Pound. 1925. Jurisprudence.

Roscoe Pound. 1932. Hierarchy of sources and forms in different systems of law. *Tul. L. Rev.*, 7:475.

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.

Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024. Chain of logic: Rule-based reasoning with large language models. *arXiv preprint arXiv:2402.10400*.

Jabez Gridley Sutherland. 1891. *Statutes and Statutory Construction: Including a Discussion of Legislative Powers, Constitutional Regulations Relative to the Forms of Legislation and to Legislative Procedure, Together with an Exposition at Length of the Principles of Interpretation and Cognate Topics*. Callaghan.

Alan Watson. 1982. Legal change: sources of law and legal culture. *U. Pa. L. Rev.*, 131:1121.

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.

Kai Yang. 2010. On the distinction between violence in robbery and theft. `http://www.jsfy.gov.cn/article/78069.html`. Accessed: 2025-02-16.

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*.

Mingkai Zhang. 2010. Justification grounds and the system of crime constitution. *The Jurist*, (1):31–39, 176–177.

Mingkai Zhang. 2017. The judicial application of the hierarchical theory. *Tsinghua Law Journal*, 11(5):20–39.

Wenxian Zhang and Wangsheng Zhou. 2007. *Jurisprudence (3rd Edition)*. Higher Education Press, Beijing.

Zhihai Zhang. 2007. On the violent elements of robbery. *Legal System and Society*, (1):222.

Guangquan Zhou. 2017a. Debate on the theory of crime constituent elements and its long-term impact. *Politics and Law*, (3):17–34.

Guangquan Zhou. 2017b. The hierarchical theory of crime and its practical development. *Tsinghua Law Journal*, 11(5):84–104.

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.

Daiming Zou. 2002. Robbery case no. 159: How to qualify the act of confined imprisonment for the purpose of robbery. *Criminal Trial Reference*, 1(24). Issue 1, Total Issue 24.

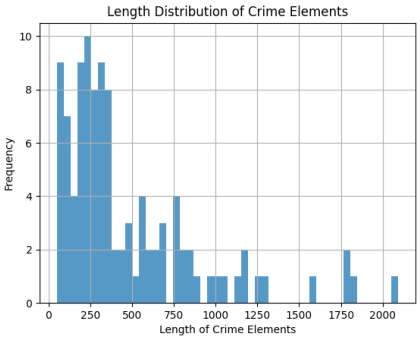

Figure 3: The average length distribution of total four elements annotated by experts.

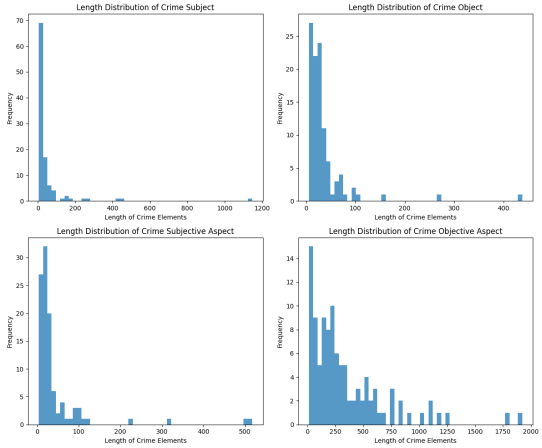

Figure 4: The length distribution of each element annotated by experts.

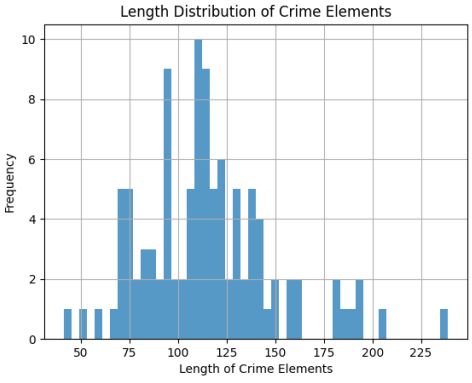

Figure 5: The average length distribution of total four elements generated by LLM.

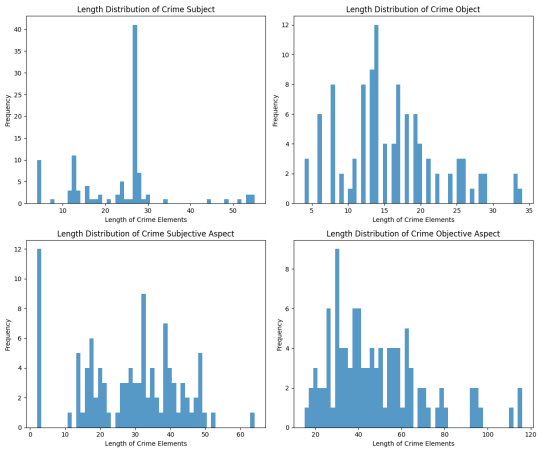

Figure 6: The length distribution of each element generated by LLM.

A Detailed Data Distribution for each Element

B Interpretation Methods

1. Literal Interpretation

A strict textual analysis method that adheres to the ordinary meaning of words as understood by a reasonable person at the time of enactment, excluding subjective intent inference

2. Systematic Interpretation

An approach interpreting legal provisions through their position within the codified legal hierarchy and logical connections with related norms, maintaining the integrity of the legal system (aligned with Dworkin's "law as integrity" theory).

3. Purposive Interpretation

A method discerning the objective legislative purpose through analysis of statutory structure and functional goals, distinct from subjective legislative intent (following Hart & Sacks' legal process school).

4. Historical Interpretation

Interpretation based on legislative history materials including drafts, debates and official commentaries, while distinguishing original meaning from framers' subjective intentions (as per Brest's original understanding theory).

5. Comparative Interpretation

A methodology referencing functionally comparable legal systems sharing common juridical traditions, employing analogical reasoning while considering local legal culture (developed through Gottfried Wilhelm Leibniz's comparative law framework).

6. Sociological Interpretation

Interpretation evaluating social efficacy through empirical analysis of implementation effects, guided by Pound's sociological jurisprudence principle that "law must be measured by its achieved results".

C Human Evaluation Guidance

The annotators included three postgraduate students specializing in criminal law 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 5) 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).

D Details for Similar Charge Disambiguation

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$_{\text{Law}}$, 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.

MALR is a up to date multi-agent framework designed to enhance complex legal reasoning (Yuan et al., 2024), enabling LLMs to autonomously decompose legal tasks and extract insights from legal rules. As its full implementation is not publicly available, we use the released code for the auto-planner module and implement the legal insight extraction following the specified steps and prompts, with necessary refinements. Experiments on the paper's reported examples show that our implementation produces task decompositions and outputs largely consistent with the original results.

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 and MALR, which require reasoning processes and prediction results, all other methods only require the output of prediction results.

E Baselines in Legal Case Retrieval

BERT(Devlin, 2018) is a language model widely used in retrieval tasks. In this paper, we chose

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

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 6: 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.

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 judg-

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

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

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

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 core factors for determining the constitution of a criminal charge.
[The basic concepts of the Four-Element Theory]
Please Compare the case facts to determine which charge's four elements they align with, thereby identifying the charge.

Table 7: Prompt template for adding the Four-Element Theory and specific four elements of crime in charge disambiguation.

Method	**GPT-4o**	**GPT-4o+Article**	**Legal-COT**	**GPT-4o+FET$_{LLM}$**	**GPT-4o+FET$_{Experts}$**
Pre-task	None	None	None	LLM-generated four elements	Expert-annotated 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**
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_{LLM}	0.4167	0.5388	0.7006	0.7925
$\text{FET}_{\text{Expert\_guided}}$	0.4201	0.5396	**0.7010**	**0.7927**

Table 9: SCR results on the full set of LeCaRDv2. Bold fonts indicate leading results in each setting. The expert-guided FET method achieved the best performance among all language models and attained the top results in both R@500 and R@1000.

ment prediction with legal document retrieval in a sequence-to-sequence manner. Since the output of GEAR cannot directly evaluate NDCG, the official results under the same setting are directly referenced in this paper. *LLM* and *Expert* represent the results of retrieval using only the four elements.

F 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.

G A Case Study of LCR

Table 10 presents a case study on the Crime of Embezzlement. By comparing the four elements annotated by experts for the crime in JUREX-4E, the case-specific four elements generated directly by the LLM, and those generated by the LLM with expert four elements of charge as guidance, we can observe that:

1) Incorporating expert fine-grained annotations enables the model to better grasp the elements of a crime, thereby providing more precise element comparison. For example, LLMs can identify the "integrity of official duties", and the subjective aspect "Intentional" can be interpreted as "having the purpose of illegally possessing public or private property", highlighting the characteristics of "official duties". Capturing the core information of the case is crucial for matching cases with similar facts.

2) LLMs can conduct case-tailored specific analysis based on the constitutive elements of a crime. Blue parts show the LLMs can better analyze the defendant's workplace and the actions taken in the case, which reflects the significance of specific and accurate legal knowledge.

Document	*[Head of document]*...In April 201X, Company A appointed B as the Sales Manager and Deputy Manager of the Catering Department, responsible for collecting outstanding debts. From December 201X to April 201Y, B embezzled X yuan from six units' payments for personal investment and consumption. In May 201Y, knowing that the company required him to reconcile accounts with the finance department, B fled after changing his phone number. It was found that after the case broke out, B's father C had returned the X yuan to Company A.		
Charge	Crime of Embezzlement		
	Experts Four-elements	**LLM-generated**	**Experts-guided**
Four Elements	**Object:** The ownership of property and integrity of duties in companies, enterprises, or other organizations. The objects include movable and immovable, tangible and intangible property of the perpetrator's organization. **Objective Aspect:** Using positional convenience to illegally appropriate organizational property in a significant amount. **Subject:** Special subject, namely personnel of companies, enterprises, or other organizations. **Subjective Aspect:**Intentional, with the purpose of illegally possessing property.	**Object:** The ownership of property of companies, enterprises, or other organizations. **Objective Aspect:** Using the convenience of one's position to illegally appropriate the property of the organization, in a significant amount. **Subject:** Personnel of companies, enterprises, or other organizations. The defendant is B. **Subjective Aspect:** Intentional.	**Object:** The ownership of property of companies, enterprises, or other organizations and the integrity of official duties. In this case, crime objetcs are the properties of the company, enterprise, or other organization where the perpetrator is employed. **Objective Aspect:** The act of using one's position for personal gain by illegally appropriating the property of the organization, where the amount involved is significant. Specifically, B used his position to embezzle X yuan from the payments of six units for personal investment and consumption. **Subject:** Special subject, namely personnel of companies, enterprises, or other organizations. Defendant B is the Sales Manager and Deputy Manager of the Catering Department of the company. **Subjective Aspect:** Intentional, with the purpose of illegally possessing public or private property.

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 **Experts-guided** indicate whether using JUREX-4E's four-elements of the crime to guide LLM in generating the four elements. Red parts mean the knowledge from JUREX-4E, while blue parts show the LLM's internal knowledge. By incorporating JUREX-4E, the model better emphasizes conviction and sentencing related information and provides more detailed descriptions of critical case facts.