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

Anonymous ACL submission

Abstract

In recent years, Large Language Models (LLMs) have been widely applied to legal tasks. To enhance their understanding of legal texts and improve reasoning accuracy, a promising approach is to incorporate legal theories. One of the most widely adopted theories is the Four-Element Theory (FET), which defines the crime constitution through four elements: Subject, Object, Subjective Aspect, and Objective Aspect. While recent work has explored prompting LLMs to follow FET, our evaluation demonstrates that LLM-generated four-elements are often incomplete and less representative, limiting their effectiveness in legal reasoning. To address these issues, we present JUREX-4E, an expert-annotated four-elements knowledge base covering 155 criminal charges. The annotations follow a progressive hierarchical framework grounded in legal source validity and incorporate diverse interpretive methods to ensure precision and authority. We evaluate JUREX-4E on the Similar Charge Distinction task and apply it to Legal Case Retrieval. 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. The dataset and code are available at: https://anonymous.4open.science/r/JUREX-86B9/

1 Introduction

Large Language Models (LLMs) have recently demonstrated impressive performance in legal tasks such as charge prediction (Yuan et al., 2024) and legal case retrieval (Feng et al., 2024). In these applications, a key challenge is accurately understanding complex legal language. To address this, recent studies have introduced legal theories into LLM workflows (Jiang and Yang, 2023; Servantez et al., 2024; Yuan et al., 2024; Deng et al., 2023), as these theories provide structured reasoning frameworks and domain knowledge. Among these theories, the Four-Element Theory (FET) in Chinese criminal law (Liang, 2017) is particularly important, as it defines the legal criteria for establishing criminal liability. FET breaks down a criminal charge into four elements: Subject, Object, Subjective Aspect, and Objective Aspect, which serve as the essential criteria for determining whether a defendant's behavior constitutes a specific crime.

Four-elements of Embezzlement

Subject	State functionaries.
Object	the management order of public funds and the integrity of officials' conduct. **Missing Object:** right to benefit from the use of public funds.
Objective Aspect	One of the following circumstances involving the misappropriation of public funds by taking advantage of one's position: 1. Misappropriating public funds *for personal use* to engage in illegal activities; ..... **Lacking explanation of** *for personal use*: specific situations such as for oneself, relatives, or others.
Subjective Aspect	Intentional.

Figure 1: An example of LLM-generated four-elements.

Most current approaches rely on the LLM's internal knowledge to incorporate the FET. A common method is to ask LLMs to emulate expert reasoning processes. For example, designing four separate prompts to guide the LLM outputs in the form of four-elements (Deng et al., 2023). This raises a critical question: Can LLMs reliably understand and apply the FET?

To investigate this, we conducted a pilot study where we provided LLMs with legal articles and asked them to generate the four elements for several representative charges (Ouyang et al., 1999). Result shows that the LLM-generated four-elements are often not accurate enough. As shown in Figure 1, in the charge of *embezzlement*, the LLM failed to identify right to benefit from the use of public funds, a core part of the Object. These results suggest that LLMs lack the domain knowledge and legal reasoning precision required for reliable FET

application.

To help LLMs better utilize the FET in legal tasks, we construct **JUREX-4E: JUR**idical **EX**pert-annotated **4-E**lement knowledge base for legal reasoning. JUREX-4E is annotated through a progressive Hierarchical Legal Interpretation System, in which legal experts annotate each element of a charge by referencing legal sources in descending order of legal validity: Criminal Articles → Judicial Interpretations → Guiding Cases → Academic Discourses. Multiple legal interpretation methods are employed to articulate the meaning, internal logic, and application cases of each element. The knowledge base covers 155 high-frequency charges, each annotated by legal experts over seven months, with an average four-element length of 472.5 words.

To assess the quality of JUREX-4E, we conducted a human evaluation grounded in a normative legal framework (Zhang, 2007a), which defines four independent dimensions: Precision, Completeness, Representativeness, and Standardization. The expert-annotated four-elements achieved an average score of 4.60 on a 5-point scale, significantly outperforming the LLM-generated ones, which scored 3.96. Among the four dimensions, the largest performance gaps appeared in Completeness and Representativeness, as expert annotations provided more comprehensive legal interpretations and summarized typical application scenarios, which are often overlooked by LLMs.

To further evaluate the quality and utility of JUREX-4E, we conducted two downstream tasks: Similar Charge Distinction (SCD) and Legal Case Retrieval (LCR). In the SCD task (Liu et al., 2021), we tested whether different charges could be more effectively distinguished by incorporating four-element knowledge. Results show that expert-annotated four-elements from JUREX-4E consistently outperformed LLM-generated counterparts across various prompting strategies and model types, improving average accuracy by 0.70% and F1-score by 0.75%. In the LCR task(Li et al., 2024d), we incorporated JUREX-4E into the retrieval pipeline to guide case-level four-element generation and similarity matching, achieving better retrieval accuracy. Together, these findings validate the high quality and practical value of JUREX-4E in enhancing legal understanding and decision-making.

Our contributions are as follows:

(1) We demonstrate that while LLMs can assist legal reasoning to some extent, they still fall short in accurately understanding and applying the Four-Element Theory.

(2) We construct the JUREX-4E, the first expert-annotated legal knowledge base grounded in a hierarchical legal interpretation framework based on legal source validity.

(3) We validate the quality and effectiveness of JUREX-4E on two representative legal tasks, SCD and LCR, where it consistently outperforms LLM-generated representations across various prompting strategies.

2 Background

The Four-Element Theory (FET) of crime constitution is a fundamental framework in Chinese criminal law (Liang, 2017). It provides a standardized structure to determine criminal liability through four elements: **Subject, Object, Subjective Aspect, and Objective Aspect**.

For the legal community, FET plays a central role in doctrinal analysis and judicial reasoning. It serves as the legal basis for both legislation and adjudication, ensuring internal consistency and normative rigor in criminal law application (Li, 2006; Zhang, 2007a). For the legal AI community, FET offers a task-agnostic and interpretable framework for modeling legal reasoning (Deng et al., 2023; Yuan et al., 2024).

For example, the four-elements of *Affray* can be briefly summarized as follows:

(1) Subject (the person who commits a criminal act and should bear criminal responsibility): Principal organizers and other active participants who have reached the age of criminal responsibility.

(2) Object (the legal interest harmed by the act): Public order.

(3) Subjective Aspect (the offender's mental state regarding the harmful act): Direct intent, where the person knowingly and willfully engages in the act of affray.

(4) Objective Aspect (the external facts of the criminal activity, including key actions and their outcomes): Acts of organizing or participating in group fighting, resulting in serious injuries.

3 Related Work

With the rise of open-source base LLMs, lots of legal LLMs have emerged, such as Lawyer LLaMA(Huang et al., 2023), ChatLaw(Cui et al., 2023), DiSC-LawLLM(Yue et al., 2023), and

TongyiFarui[1]. These models are typically adapted from general-purpose LLMs via domain-specific post-training or Retrieval-Augmented Generation (RAG), incorporating legal texts like cases and laws.

Although these models achieve notable improvements on legal tasks, they still struggle with complex legal reasoning, such as distinguishing similar charges or excluding distracting case details(Hu et al., 2025). To further enhance model performance, particularly in tasks requiring complex legal reasoning, some studies draw inspiration from established legal reasoning paradigms. For example, introducing the legal syllogism for legal judgment prediction(Jiang and Yang, 2023); using the IRAC paradigm to guide LLMs in reasoning about compositional rules(Servantez et al., 2024). Several works have drawn on the FET in the context of Chinese criminal law. For example, breaking down legal rules into FET-aligned components using automated planning techniques (Yuan et al., 2024); employing model-generated FETs 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 Four-Element Theory, without systematically validating this assumption.

4 Can LLM Grasp Legal Theory?

To examine whether LLMs can internalize and apply the Four-Element Theory (FET), we asked LLMs to generate the four-elements for several representative charges. This task reflects whether the model can use its internal knowledge to analyze the FET.

We selected GPT-4o as the evaluation LLM, as it achieves state-of-the-art performance on legal benchmarks (Fei et al., 2023; Li et al., 2024c) and in our pilot study (Appendix D), indicating a strong capacity to understand and apply legal knowledge. Following prior work(Deng et al., 2023; Cui et al., 2024; Zhou et al., 2023), for each charge, we prompted GPT-4o with corresponding criminal articles(see prompt template in Appendix C).

We invited legal experts who passed the bar exam to analyze the LLM-generated Four-Element Theories (FETs) and identified two main issues:

(1) Inaccurate elements: LLMs may produce inaccurate four-elements. For example, in Figure 1, for *Embezzlement*, the LLM-generated Object is "the management order of public funds and the integrity of officials' conduct", missing the right to benefit from the use of public funds, which is necessary to identify this charge.

(2) Lack of interpretive awareness: LLMs fail to recognize when statutory language requires deeper interpretation. In the same case, the model simply extracts "misappropriating public funds for personal use." to describe the Objective Aspect. However, this phrase is far too general for practice. In judicial interpretations[2], the term "for personal use" should be interpreted with three situations: (1) using public funds for oneself, relatives, or other individuals; (2) lending public funds to other entities in one's own name; or (3) using public funds in the name of one's organization for another entity to gain personal benefits.

The lack of legal grounding in current LLMs leads to inaccurate generation of the four essential elements of a crime (FETs), which in turn undermines the reliability of legal reasoning tasks. To overcome this, we introduce an expert-annotated FET dataset that captures both formal legal definitions and practical interpretive nuances, supporting more trustworthy and adaptable legal AI systems.

5 Dataset Construction

To ensure both legal validity and interpretive clarity, we design a hierarchical annotation framework rooted in statutory sources and authoritative interpretive methods.

5.1 Hierarchical Legal Interpretation System

Given a specific charge, we ask legal experts to annotate the four-elements based on relevant legal materials like articles and cases. This annotation process is essentially an act of legal interpretation. *Legal interpretation* refers to the application of various methods to analyze and understand legal texts, to determine their meaning and application in specific legal contexts(wha, 2005). In our task, it involves applying different interpretation methods to the different materials in order to analyze and define the connotation and extension of each of the four-elements of a charge. In designing our

[1] https://tongyi.aliyun.com/farui

[2] National People's Congress Standing Committee. *Interpretation on Article 384, Paragraph 1 of the Criminal Law of the People's Republic of China*, adopted at the 27th Meeting of the Standing Committee of the 9th National People's Congress on April 28, 2002.

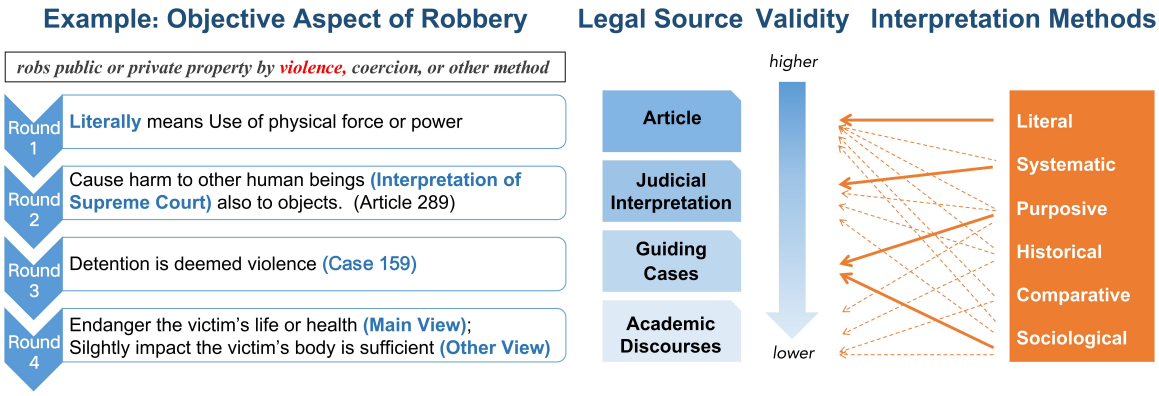

Figure 2: Hierarchical Legal Interpretation System based on legal source validity. The system consists of four annotation rounds, each using different interpretive methods based on different legal sources. Solid arrows indicate the primary method applied; dashed arrows represent supplementary use.

annotation framework, we address the following two questions:

(1) **What sources are interpreted.** Legal interpretation draws upon various legal texts with different levels of validity. In legal studies, these sources are categorized based on their legal validity into formal sources (which carry legal force in judgments) and informal sources (which serve as references without legal force)(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). Accordingly, we organize legal sources by their level of validity, with the following order of priority: Article → Judicial Interpretations → Guiding Cases → Academic Discourses.

(2) **How the law is interpreted.** When interpreting the above sources, different interpretation methods are required. These methods follow a hierarchical order(Sutherland, 1891; Kim and Division, 2008; Eig Larry, 2014): Legal interpretation should begin with literal interpretation (interpreting the text based on its plain meaning). If the intended meaning cannot be clearly derived from the article alone, systematic interpretation (considering the article's role within the legal system) and purposive interpretation (considering the legislative intent) should be applied. If ambiguity remains, historical interpretation (based on the legislative history) and sociological interpretation (based on the article's social function and consequences) may be used to further clarify the legal meaning. The specific definition of legal interpretation methods is in Appendix B.

We also consider the nature of each source. For example, as an informal source, guiding cases can't define an element literally but can supplement it through purposive and sociological interpretation. The correspondence between interpretation methods and legal sources is illustrated by the arrows in Figure 2.

5.2 Annotation Process

As shown in Figure 2, our annotation process takes charges as input and outputs corresponding four-elements, following a Hierarchical Legal Interpretation System to organize legal sources by validity and apply interpretation methods. Annotators are experts recruited from Chinese colleges, all of whom have passed the bar exam and are familiar with FET. The entire annotation process took 7 months and involved four rounds.

The First Stage: Article. The interpretation of the four-elements starts from the charge's corresponding articles with the highest legal validity, mainly through *textual interpretation*.

At this level, annotators analyzed each article's subject–predicate–object structure to identify candidate elements. For example, in the article concerning *robbery*, the phrase 'forcibly seizing public or private property through violence, coercion, or other means' describes the Objective Aspect. Since the article lacks an explicit subject, a general subject is assumed by default. The adverbs 'violence' and 'coercion' indicate an intentional act.

To ensure consistency across different charges, legal terms are categorized and standardized. For instance, subjective aspects are classified as either intentional or negligent.

The first stage spans two months and serves as

the foundation of the annotation process. When literal interpretation fails to clearly explain an element, further interpretive methods are applied in the subsequent level of annotation.

The Second Stage: Judicial Interpretation. In this stage, we refine legal elements using judicial interpretations. The primary method at this level is systematic interpretation, which places the article corresponding to an ambiguous element within the broader legal context. By examining the article's structural role and its connections to related articles and judicial interpretations, annotators further determine the precise meaning and scope of the element. This stage also spans two months.

For example, to clarify whether 'violence' in robbery must target persons or may include property, systematic interpretation refers to Article 289(Congress, 2017) in Chinese criminal law, which shows that violence against property can also constitute robbery, highlighting how legal context resolves ambiguity.

The Third Stage: Guiding Cases. Although the first two levels define the four-elements based on articles and judicial interpretations, their application in concrete cases often demands further interpretation. In practice, Guiding Cases, designated by the Supreme People's Court since 2011 as references for adjudicating similar disputes(Chen et al., 2024), play a crucial role in resolving such ambiguities.

Therefore, at the third level, we focus on disputed aspects of the four-elements across charges and refine them through representative Guiding Cases. Specifically, we apply *purposive and sociological interpretation* to examine how the elements are interpreted in judicial practice, considering both legislative intent and social context. These methods bridge the subtle gap between abstract legal theory and practical cases. This stage spans one month.

For example, in defining 'violence' in robbery, purposive and sociological interpretations clarify whether acts such as illegal detention(Zou, 2002) or molestation(Ma, 2021), though not physically injurious, can suppress resistance and thus qualify as violent means. Such cases guide the inclusion of these acts as valid objective aspects in robbery annotations.

The Fourth Stage: Academic Discourses. To ensure the extensibility and academic depth of the dataset, we incorporate academic discourses to further refine elements that remain widely contested. We apply interpretive methods such as *comparative interpretation, purposive interpretation, and sociological interpretation*. These methods help contextualize controversial elements by referencing debates in legal scholarship. We highlight key areas of disagreement, distinguish between mainstream and minority views, and provide concise annotations explaining the underlying legal reasoning. This stage supplements the dataset with richer legal perspectives and supports future adaptation to evolving academic and judicial interpretations. This stage spans one month.

For example, in defining 'violence' in robbery, mainstream views in China, the former Soviet Union, North Korea, and Japan require that it endanger the victim's life or health (Zhang, 2007b), while others argue that any force sufficient to subdue the victim should qualify (Yang, 2010). Our annotations mark both the dominant consensus and minority positions.

5.3 Data Distribution

Metric	LLM_{Mean}	$\text{LLM}_{\text{Median}}$	$\text{Expert}_{\text{Mean}}$	$\text{Expert}_{\text{Median}}$
Avg. Length	115.43	-	472.53	-
Subject	23.12	27	51.64	17
Object	15.86	15	36.01	25
Subjective Aspect	28.00	30	42.38	21
Objective Aspect	48.45	45	342.5	230

Table 1: Comparison of element lengths.

Our dataset includes 155 criminal charges, selected based on their frequency in practice (see Appendix A).

To compare the quality of expert FETs and LLM-generated FETs, we selected 105 charges in JUREX-4E that also appear in the widely used Lecard-V2 dataset (Li et al., 2024d). LLM-generated FETs were produced using the same setup as before, with a generation context of 8192 tokens. Table 1 summarizes the differences in element length, with full distributions available in Appendix A.

Overall, expert FETs are significantly longer, with an average total length of 472.53 tokens compared to 115.43 for LLM-generated ones. The most pronounced gap appears in the Objective Aspect (OA)(mean: 342.5 vs. 48.45), where experts provide detailed factual descriptions, such as action, result, time, and location, often underdeveloped in LLM outputs. While the Subject (SB), Object (OB), and Subjective Aspect (SA) show smaller median differences, notable variation remains, especially in SB (mean: 51.64 vs. 17), which in cer-

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.

tain charges involves complex legal interpretations (e.g., "work" in copyright infringement) requiring more elaborate legal definitions.

6 Human Evaluation

To compare the quality of expert-annotated and LLM-generated FETs, we selected six complicated charges in Chinese judicial practice (Ouyang et al., 1999). Based on prior theoretical framework (Zhang, 2007a), we assess the quality of FETs along four independent dimensions: **Precision, Completeness, Representativeness, and Standardization**.

- Precision: Evaluate whether each element accurately aligns with its statutory definition, reflecting key terms in the corresponding legal article.
- Completeness: Evaluates whether each element includes all practically necessary information, ensuring the definition is sufficient to guide legal reasoning.
- Representativeness: Evaluates whether the annotations reflect the most typical and practically significant scenarios in judicial practice, such as common forms of harm in intentional injury cases.
- Standardization: Evaluates whether the expressions of elements are consistent across different charges, with clear, concise, and unambiguous language that facilitates understanding and minimizes interpretive variance.

Each dimension was scored by experts from two backgrounds: one group with a purely legal background and another with a combined background in law and AI, 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 E.

As shown in Table 2, expert annotations consistently outperform LLM-generated elements across all four dimensions. The most pronounced deficiencies are observed in Completeness (+0.86) and Representativeness (+0.88). This aligns with our earlier analyses, where expert-generated elements include more factual details and representative descriptions. The gap in Precision (+0.57) suggests a tendency toward vague or legally irrelevant content, while the smaller difference in Standardization (+0.23) shows that LLMs can mimic structural patterns but lack deeper normative consistency. These results reinforce the importance of expert supervision in providing reliable legal knowledge.

7 Evaluation on Similar Charge Disambiguation

To further validate annotation quality, we introduce the Similar Charge Disambiguation (SCD) task(Yuan et al., 2024; Li et al., 2024a). Given the case fact and a set of similar charges, SCD task requires the model to identify which charge is correct. We evaluate whether similar charges can be effectively distinguished based on their four-elements, and whether expert-annotated FETs perform better than LLM-generated FETs.

7.1 Experiment Settings

7.1.1 Dataset and metrics

We chose the SCD dataset released by (Liu et al., 2021). Following previous work(Yuan et al., 2024), we selected three 2-label classification groups: Fraud & Extortion (F&E), Embezzlement & Misappropriation of Public Funds (E&MPF), and Abuse of Power & Dereliction of Duty (AP&DD). Each charge has over 1.9k cases, with a total of 13,962 cases. The details of the groups are shown in Appendix F. Following previous work (Liu et al., 2021; Yuan et al., 2024), we use Average Accuracy (Acc) and macro-F1 (F1) as evaluation metrics.

7.1.2 Baselines and Methods

We compared the following baselines: **GPT-4o (Achiam et al., 2023)** and **GPT-4o+Article**, which explicitly supplies relevant legal articles; **Legal-CoT (Kojima et al., 2022)**, a Chain-of-Thought variant that applies the Four-Element Theory step by step, and **MALR (Yuan et al., 2024)**, a multi-agent framework that decomposes legal tasks into FET-aligned subtasks. Details are in Appendix F.

Methods: Following Section 4, our main model is GPT-4o. We also compared Farui-plus (the latest

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
Farui-plus+FET_{4o}	89.09	90.27	86.32	88.00	75.90	77.67	83.77	85.31
Farui-plus+$\text{FET}_{\text{Expert}}$	89.29	90.98	86.13	87.54	76.25	78.12	83.89	85.55
Qwen2.5-72b+FET_{4o}	93.15	95.06	90.99	93.56	87.71	88.56	90.62	92.39
Qwen2.5-72b+$\text{FET}_{\text{Expert}}$	93.29	95.18	91.18	93.66	87.81	89.45	90.76	92.76
GPT-4o+$\text{FET}_{\text{farui}}$	94.86	96.12	91.84	92.64	89.35	89.85	92.02	92.87
GPT-4o+FET_{qwen}	95.53	96.53	91.82	92.96	89.48	90.09	92.28	93.19
GPT-4o+$\text{FET}_{\text{4o+farui+qwen}}$	94.97	96.24	91.84	92.73	89.69	90.12	92.17	93.03
GPT-4o+FET_{4o}	95.73	96.56	91.87	92.01	89.61	89.69	92.40	92.75
GPT-4o+$\text{FET}_{\text{Expert}}$	**96.06**	**96.69**	92.57	**93.05**	**90.53**	**90.62**	**93.05**	**93.45**

Table 3: Performance on the Similar Charge Disambiguation (SCD) task. "Expert" refers to out expert-annotated FET, while "4o", "qwen", and "farui" refer to FET generated by different LLMs. Highest results are in bold.

version of Tongyifarui, representative legal LLM) and Qwen2.5-72B(Bai et al., 2023) (representative open-source LLM). To incorporate FET knowledge, each group of similar charges is augmented with four-element descriptions, either generated by LLMs or sourced from JUREX-4E. For example, GPT-4o+FET_{LLM} uses LLM-generated FETs, while GPT-4o+$\text{FET}_{\text{Expert}}$ uses expert-annotated ones. The input format is fixed across methods, differing only in the *[four-elements of candidate charges]* (Appendix F). All experiments are zero-shot, with max_tokens set to 3,000 (10,000 for Legal-CoT and MALR) and a temperature of 0 or 0.0001 in repeated runs.

7.2 Results

The SCD results are shown in Table 3, where we can observe that:

Effectiveness of Structured FET Knowledge: Providing specific structured charge FETs yields the highest accuracy among all legal knowledge integration methods. Compared to implicit approaches, such as prompts (GPT-4o+Article, Acc 92.09) or reasoning chains (Legal-COT, Acc 89.95), structured FET knowledge offers more effective support for legal decision-making(e.g., GPT-4o+$\text{FET}_{\text{Expert}}$, Acc 93.05)

Superiority of Expert-Annotated FET: Expert-annotated FET consistently outperforms LLM-generated FET across three representative LLMs, including $\text{FET}_{\text{farui}}$, FET_{qwen}, FET_{4o}, and their combination ($\text{FET}_{\text{4o+farui+qwen}}$). For example, GPT-4o+$\text{FET}_{\text{Expert}}$ surpasses GPT-4o+FET_{4o} by 0.65 in average accuracy and 0.70 in F1-score.

Consistent Gains Across Models: Expert-annotated FETs yield consistent performance gains across different SCD models. When applied to Farui-plus, Qwen2.5-72b and GPT-4o, it improves F1-score by +0.24, +0.37, and +0.70 respectively over their LLM-generated FET baselines.

8 Application in Legal Case Retrieval

In this section, we design a simple expert-guided FET method to apply JUREX-4E to Legal Case Retrieval (LCR) task, which retrieves relevant cases based on case facts. This task is well-suited for FET because it requires a comprehensive comparison of the four-elements across different charges in cases.

8.1 Dataset and Metrics

LeCaRDv2(Li et al., 2024d) is the latest version of LeCaRD(Ma et al., 2021), which is widely used in legal 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. Following previous work (Qin et al., 2024), we chose 1390 candidates and used NDCG@10, 20, 30, Recall@1, 5, 10, 20, and MRR as metrics. We also tested different candidate pool settings (see Appendix G). The result is consistent.

8.2 Baselines and Methods

We adopt a dense retrieval framework based on BGE-m3 (Chen et al., 2023), a strong embedding model for legal and general-domain texts. Given a query q and a candidate case c, we compute their vector representations using a shared BGE-m3 encoder. Retrieval is performed by computing cosine similarities between the query and all candidates and selecting the top-k candidates.

To enhance retrieval accuracy, we compare the following three methods:

Model	NDCG@10	NDCG@20	NDCG@30	R@1	R@5	R@10	R@20	R@30	MRR
BGE (case_fact only)	0.4737	0.5539	0.5937	0.0793	0.2945	0.4298	0.6500	0.7394	0.1926
BGE+FET (Qwen2.5)	0.5125	0.5858	0.6350	0.1104	0.2870	0.4653	0.6679	0.7836	0.2168
FET only	0.3367	0.3971	0.4487	0.0622	0.2006	0.3279	0.4806	0.6037	0.1524
BGE+FET (Expert, Qwen2.5)	**0.5295**	**0.5979**	**0.6416**	**0.1124**	**0.3122**	0.4838	0.6791	0.7824	**0.2206**
FET only	0.3354	0.4035	0.4541	0.0849	0.1923	0.3076	0.4839	0.6097	0.1606
BGE+FET (GPT-4o)	0.5139	0.5862	0.6291	0.0980	0.2967	0.4769	0.6802	0.7828	0.2140
FET only	0.3583	0.4293	0.4798	0.0506	0.2240	0.3644	0.5383	0.6652	0.1453
BGE+FET (Expert, GPT-4o)	0.5211	0.5920	0.6379	0.1024	0.3049	**0.4883**	**0.6885**	**0.7967**	0.2155
FET only	0.3766	0.4584	0.5111	0.0715	0.1894	0.3709	0.5891	0.7203	0.1624

Table 4: Performance on the Legal Charge Retrieval (LCR) task. The highest results are in bold. "FET only" indicates using the four-element descriptions without case facts.

(1) **BGE(case_fact only)**: Standard dense retrieval using only BGE-m3 embeddings of the raw case facts.

(2) **BGE+FET ($\mathcal{M}_g$)**: We prompt different LLMs $\mathcal{M}_g$ to generate a structured four-element description of each case (case-FET) based solely on its facts, without using external knowledge. These case-FETs are then embedded with BGE-m3, and used to compute similarity. Because the FET abstracts away case-specific details, we combine the original fact-based similarity and the FET-based similarity in a ratio of 7:3.

(3) **BGE+FET (Expert, $\mathcal{M}_g$)**: A expert-guided FET method that incorporates JUREX-4E to guide case-FET generation. It consists of four steps:

1. Charge Prediction. A charge prediction model $\mathcal{M}_p$ (Qwen-plus, details see Appendix D) predicts the set of likely charges $Z = \{z_1, ..., z_k\}$ for the query case.
2. Expert FET Matching. Retrieving corresponding charge's four-elements $\{f_z\}_{z \in Z}$ for each predicted charge in JUREX-4E. These provide theoretical guidance for subsequent reasoning.
3. Case-FET Generation. Guided by $\{f_z\}$, the LLM $\mathcal{M}_g$ generates case-specific four-elements fet_c for candidate c.
4. Dense retrieval. We embed the generated FETs using BGE-m3 and compute similarity scores as in Method (2), combining both factual and FET-based similarities.

For the $\mathcal{M}_g$, we chose Qwen2.5-72b and GPT-4o. The retrieval framework is implemented with the FlagEmbedding Toolkit[3] with an RTX 3090. Following prior work(Li et al., 2024d; Qin et al., 2024), we also compare some dense retrieval methods to examine the representativeness of BGE-m3. Results of baselines and prompt templates are available in Appendix G.

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

8.3 Results

The LCR results are shown in Table 4, where we can observe that: **(1) FET Enhances Retrieval.** Integrating the FET improves retrieval performance across all metrics. For instance, BGE+FET(GPT-4o) improves MRR by 11.11%, and BGE+FET (Expert, GPT-4o) achieves an even larger gain of 11.89%, indicating that structured legal theory benefits retrieval quality. **(2) Expert Knowledge is Important.** Expert-guided case-FET consistently outperforms LLM-generated variants across both Qwen2.5-72b and GPT-4o backbones. For example, BGE+FET (Expert, GPT-4o) achieves higher Recall@30 (0.7967 vs. 0.7828) and MRR (0.2155 vs. 0.2140). The gap is even larger in the *FET only* setting(e.g., MRR 0.1624 vs. 0.1453 for GPT-4o), demonstrating that expert knowledge captures critical legal reasoning that LLMs may overlook.

We provide a case study in Appendix I. It illustrates that the expert-annotated four-elements of charges provide practical judgment points and key narratives (e.g., the special subject of *Embezzlement*) that help the LLM focus on essential facts to analyze the case-FET.

9 Conclusion

This paper presents JUREX-4E, an expert-annotated FET knowledge base built through a structured legal interpretation process and validated on downstream tasks. Grounded in widely accepted legal interpretative methods across jurisdictions and fields, our framework is adaptable to other domains. Additionally, our exploration of incorporating domain knowledge also offers insights for fields like medicine and industry, where expert knowledge is critical for decision-making.

10 Ethical Considerations

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

11 Limitations

Our current knowledge base is limited to 155 charges under Chinese Criminal Law due to the high cost of expert annotation. Future work will explore extending it to other legal domains and jurisdictions.

Another limitation lies in our current integration of factual and legal information. In the LCR task, although case facts are used to generate FETs, the *FET only* variant excludes the original case facts during retrieval, resulting in performance loss (e.g., MRR 0.1624 vs. 0.2155). This suggests that our current method remains coarse-grained, and more fine-grained fusion strategies, such as multi-agent coordination or retrieval-time integration, deserve future exploration.

Acknowledgments

References

2005. *Chapter One. What Is Legal Interpretation?*, pages 3–60. Princeton University Press, Princeton.

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*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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*.

Benjamin M Chen, Zhiyu Li, David Cai, and Elliott Ash. 2024. Detecting the influence of the chinese guiding cases: a text reuse approach. *Artificial Intelligence and Law*, 32(2):463–486.

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.

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).

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

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.

Yiran Hu, Huanghai Liu, Qingjing Chen, Ning Zheng, Chong Wang, Yun Liu, Charles LA Clarke, and Weixing Shen. 2025. J&h: Evaluating the robustness of large language models under knowledge-injection attacks in legal domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28106–28115.

Quzhe Huang, Mingxu Tao, Chen Zhang, Zhenwei An, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

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, You Chen, Qingyao Ai, Yueyue Wu, Ruizhe Zhang, and Yiqun Liu. 2024c. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *arXiv preprint arXiv:2409.20288*.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024d. 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.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

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*.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Mingkai Zhang. 2007a. Normative elements of the constitutive requirements. *Studies in Law*, (06):76–93.

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

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

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.

A Charge Selection and Detailed Data Distribution

Charge Selection: To systematically determine charge frequency, we analyzed the CAIL2018 dataset(Xiao et al., 2018), which contains 2,676,075 criminal cases annotated with 183 criminal law articles and 202 criminal charges and a total of 3,010,000 criminal charges. Apart from few charges that have been merged or changed name, our dataset largely covers all criminal charges from CAIL2018 that have a frequency of over 3,000 (>0.099%) occurrences.

Length distribution for each element: Table 6.

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 articles 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 Prompt for LLM-generated FET

See Table 5.

D Details about Pilot Study

We selected candidate models from LawBench(Fei et al., 2023) and LexEval(Li et al., 2024c), which contain the broadest and most up-to-date evaluation of legal LLMs. From these, we chose top-performing models such as GPT-4, Qwen-14B-chat, and representative legal-specific LLMs.

For best performance, we used GPT-4o(the latest version of GPT-4 at that time) and Qwen-plus(a stronger commercial variant of Qwen-2.5-72B. (Aliyun model-studio official site))

During implementation, we found that most legal LLMs were unavailable. The only stably accessible one was Farui (A leading legal LLM built on Qwen, (Aliyun model-studio official site, Tongyi Farui)), specifically the version "tongyifarui-890" from its official API.

To compare GPT-4o, Qwen-plus, and tongyifarui-890, We sampled 300 cases from our legal retrieval dataset and asked models to perform charge prediction, which is the pre-task for generating Case-FETs.

(For each case in legal retrieval, the model was required to predict charges, so we can match charges' Expert-FETs, and use them to generate Case-FET.)

This task involved all criminal charges, including multi-defendant and multi-charge scenarios, and requires models to predict charges from open text without a predefined list, making it a challenging legal task.

The result showed that GPT-4o (59.78%) > Qwen-plus (58.70%) » tongyifarui-890 (21%). Given Farui's poor performance, we did not include it in subsequent experiments.

We further evaluated GPT-4o and Qwen-plus based on their ability to generate Case-FETs. The

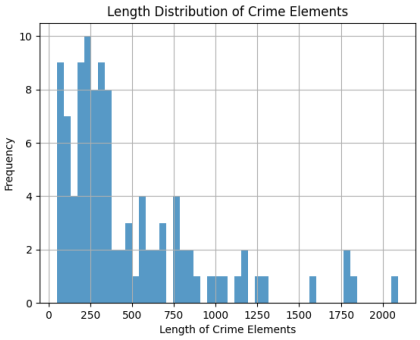

Figure 3: The average length distribution of total four-elements annotated by experts.

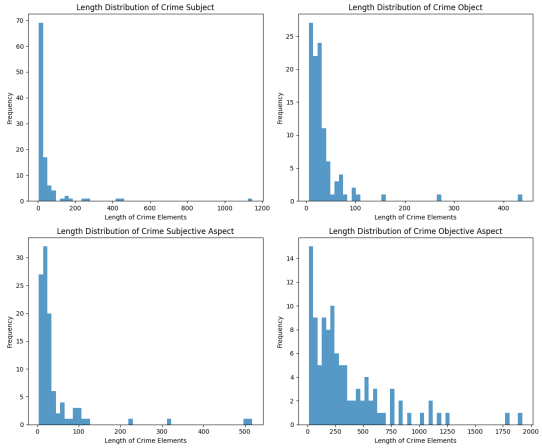

Figure 4: The length distribution of each element annotated by experts.

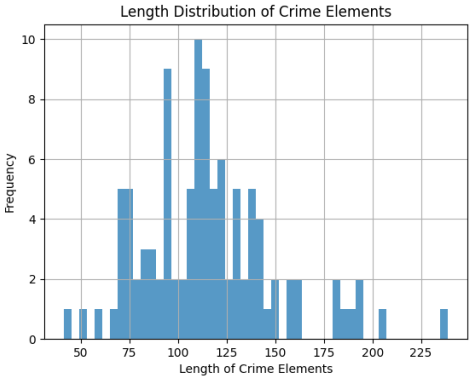

Figure 5: The average length distribution of total four-elements generated by LLM.

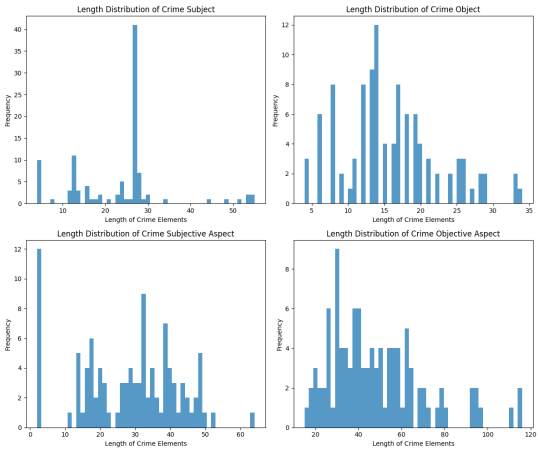

Figure 6: The length distribution of each element generated by LLM.

You are an expert in criminal law. Based on the given charge, please analyze it according to China's criminal law and output the four-elements of the charge in order, including:
- **Object**: The concretization of a certain abstract social interest. For example, the object of charges that infringe on personal rights is the right to life, while the object of property-related charges could be items such as mobile phones or wallets.
- **Objective Aspect**: The objective facts of the criminal act, including the key actions that trigger the charge (e.g., theft, robbery) and the consequences caused by the act (e.g., serious injury, death, property loss).
- **Subject**: Typically, the general subject of the charge, but in some cases, a specific subject is required (e.g., government officials in certain offenses).
- **Subjective Aspect**: The mental state of the perpetrator, such as intent or negligence.
Relevant Legal Articles: []
Please synthesize the above information to generate a refined set of four-elements that represent the characteristics of the charge.
Output format: { "Crime": "", "Four Elements of the Crime": { "Crime Object": "", "Objective Aspect": "", "Subject": "", "Subjective Aspect": "" } }
Crime: []

Table 5: Prompt template for generating the Four Elements of a Crime (FET) using LLMs

results showed that GPT-4o outperformed Qwen-plus (MRR 0.2140 v.s. 0.2052). Considering both results, we adopted GPT-4o as our primary model in the paper.

Subsequently, in efforts to improve charge prediction for matching charges' Expert-FETs, we found that Qwen-plus performed better than GPT-4o when a charge list was provided (58.70%->80.43% vs. 59.78%->71.74%). Therefore, in this specific setting for charge prediction before retrival, we used Qwen-plus.

For fair and reproducible presentation of results on specific downstream tasks (SCD and LCR), as mentioned in the main text, we present the results of open source Qwen2.5-72b.

E 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 6) 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).

F 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,

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 6: 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 7: 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.

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 9, 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.

G Baselines in Legal Case Retrieval

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

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 8: 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 9: Prompts of different methods in Similar Charge Disambiguation. # represents a format input.

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-Text2Vec	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
BGE (case_fact only)	0.4737	0.5539	0.5937	0.0793	0.2945	0.4298	0.6500	0.7394	0.1926
BGE+FET (Qwen2.5)	0.5125	0.5858	0.6350	0.1104	0.2870	0.4653	0.6679	0.7836	0.2168
FET only	0.3367	0.3971	0.4487	0.0622	0.2006	0.3279	0.4806	0.6037	0.1524
BGE+FET (Expert, Qwen2.5)	**0.5295**	**0.5979**	**0.6416**	**0.1124**	**0.3122**	0.4838	0.6791	0.7824	**0.2206**
FET only	0.3354	0.4035	0.4541	0.0849	0.1923	0.3076	0.4839	0.6097	0.1606
BGE+FET (GPT-4o)	0.5139	0.5862	0.6291	0.0980	0.2967	0.4769	0.6802	0.7828	0.2140
FET only	0.3583	0.4293	0.4798	0.0506	0.2240	0.3644	0.5383	0.6652	0.1453
BGE+FET (Expert, GPT-4o)	0.5211	0.5920	0.6379	0.1024	0.3049	**0.4883**	**0.6885**	**0.7967**	0.2155
FET only	0.3766	0.4584	0.5111	0.0715	0.1894	0.3709	0.5891	0.7203	0.1624

Table 10: Performance on the Legal Charge Retrieval (LCR) task with baselines. Highest results are in bold. "FET only" indicates using the four-element descriptions without case facts.

Prompt 1: Charge Prediction

You are a legal expert specializing in criminal law. Based on the provided list of charges, determine which charges are applicable to the given case facts. Please note that you should only output the charge names, without any additional information. The charges must be selected from the provided list and should be separated by commas.

[Crime List]
[Case Facts]

Table 11: Prompt used for charge prediction.

BERT-base-Chinese[4]. **Legal-BERT**[5](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**[6](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.

Baseline results are provided in Table 10.

To support reproducibility, we provide the full prompt templates used in our pipeline. Table 11 shows the prompt for charge prediction, and Table 12 presents the prompt used for generating four-element annotations in both BGE+FET(LLM) and BGE+FET(Expert, LLM).

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

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

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

H SCR results on the full LeCaRDv2 Dataset

As presented in Table 13, 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.

I A Case Study of LCR

Table 14 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.

Prompt 2: BGE+FET(LLM) and BGE+FET(Expert, LLM).

You are a legal expert specializing in criminal law. Based on Chinese criminal law knowledge, analyze the following case facts and provide the following information in sequence:
1. The four-elements of the crime:
 - Criminal Object: The tangible or intangible interests being infringed upon (e.g., personal rights such as life, or property rights such as money, vehicles).
 - Objective Aspect: The objective facts of the criminal activity, including key actions (e.g., theft, robbery) and consequences (e.g., injury, death, loss).
 - Criminal Subject: Typically general subjects; special subjects in certain crimes (e.g., government officials).
 - Subjective Aspect: Whether the act was intentional or negligent.
2. Charge: Only output the specific crime name(s).
3. Relevant Legal Articles: Only output the article number(s) of the relevant laws.
Output format: JSON. For each crime involved in the case, provide a separate dictionary entry.
[Output Sample]

```
{
 "Crime 1": {
  "Four Elements": {
   "Criminal Object": "Personal rights: the victim Wang's right to life; Property rights: vehicle.",
   "Objective Aspect": "The defendant Wu drove under the influence and collided with the victim Wang, causing Wang's immediate death and vehicle damage.",
   "Criminal Subject": "Defendant Wu, the driver.",
   "Subjective Aspect": "Negligence"
  },
  "Charge": "Traffic Accident Crime",
  "Relevant Legal Article": "Article 133"
 },
 "Crime 2": {
  "Four Elements": {
   "Criminal Object": "Social management order: infringement on the state's document management system; Property rights: forged documents and related items.",
    "Objective Aspect": "Defendant 1 purchased equipment and materials to forge documents. Defendant 2 delivered the forged documents. Defendant 3 facilitated transactions via the internet, handling payments and document transfers.",
   "Criminal Subject": "Multiple defendants, all individuals with full criminal responsibility.",
   "Subjective Aspect": "Intentional"
  },
  "Charge": "Forgery, Alteration, or Sale of Official Documents, Certificates, and Seals of State Organs",
  "Relevant Legal Article": "Article 280, Paragraph 1"
 }
} }
```

Table 12: Prompt for generating four-element annotations used in FET_{LLM} and $\text{FET}_{\text{Expert\_Guided}}$.

Model	**R@100**	**R@200**	**R@500**	**R@1000**
Legal-BERT	0.1116	0.1493	0.2174	0.2819
Lawformer	0.2432	0.304	0.4054	0.4833
ChatLaw-Text2Vec	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
BGE+FET(GPT-4o)	0.4167	0.5388	0.7006	0.7925
BGE+FET(Expert, GPT-4o)	**0.4201**	**0.5396**	**0.7010**	**0.7927**

Table 13: 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.

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 14: 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.