Through the MUD:
A Multi-Defendant Charge Prediction Benchmark with Linked Crime Elements

Xiao Wei[1], **Xu Qi**[1], **Hang Yu**[1*], **Qian Liu**[2*], **Erik Cambria**[3]
[1]School of Computer Engineering and Science, Shanghai University, China
[2]School of Computer Science, University of Auckland, New Zealand
[3]College of Computing and Data Science, Nanyang Technological University, Singapore
{xwei,welch,yuhang}@shu.edu.cn, Liu.Qian@auckland.ac.nz, cambria@ntu.edu.sg

Abstract

The current charge prediction datasets mostly focus on single-defendant criminal cases. However, real-world criminal cases usually involve multiple defendants whose criminal facts are intertwined. In an early attempt to fill this gap, we introduce a new benchmark that encompasses legal cases involving multiple defendants, where each defendant is labeled with a charge and four types of crime elements, *i.e., Object Element*, *Objective Element*, *Subject Element*, and *Subjective Element*. Based on the dataset, we further develop an interpretable model called EJudge that incorporates crime elements and legal rules to infer charges. We observe that predicting crime charges while providing corresponding rationales benefits the interpretable AI system. Extensive experiments show that EJudge significantly surpasses state-of-the-art methods, which verify the importance of crime elements and legal rules in multi-defendant charge prediction. Source code and dataset available at https://github.com/welchxu/MCP.

1 Introduction

The charge prediction task aims to automatically recommend charges given a fact description (Luo et al., 2017; Nair and Modani, 2023). It has attracted substantial attention recently, leading researchers to construct high-quality datasets for its advancement, such as CAIL2018 (Xiao et al., 2018) and ECHR (Medvedeva et al., 2018).

Commonly, existing datasets mainly support coarse-grained prediction, recommending charges for each defendant based on the whole criminal facts, without specifying relevant details. For example, Fig. 1 (a) shows a case from CAIL2018 (Xiao et al., 2018) that has only one defendant and is labeled solely with the charge, without any justification or rationale for the conviction.

(a) Example in previous dataset (CAIL2018)	
事实描述：11月5日上午…被告人胡某用木制坐垫打伤被害人孙某左腹部…孙某的左腹部损伤已达重伤二级。*#Translation (**Fact Description:** On the morning of 5 November… the defendant Hu injured the victim Sun's left abdomen with a wooden cushion…Sun's left abdominal injury has reached the second degree of serious injury.)*	
Defendant:	胡某 *(Hu)*
Charge:	故意伤害罪 *(Intentional Injury)*

(b) Example in our benchmark (MUD)		
事实描述:…, 刘某因邻里纠纷…后厮打在一起，在厮打过程中，刘某用尖刀将于某扎伤。致使于某股动脉破裂失血性休克死亡…, 案发后，被告人王某在明知刘某故意伤害他人的情况下，仍帮助其逃跑，致使刘某逃避法律制裁长达15年。*#Translation (**Fact Description**: …due to a dispute, the defendant Liu fought with Yu Mou, and stabbed Yu with a knife, caused Yu femoral artery rupture shock death, After the incident, Wang known Liu injure others, but still help him escape, resulting in Liu to evade justice for 15 years.*		
Defendant:	刘某 *(Liu)*	王某 *(Wang)*
Charge:	故意伤害罪 *(Intentional Injury)*	包庇罪 *(Harboring)*
Crime Elements:	**Subject Element:** 刘某 *(Liu)* **Object Element:** 公民的人身、民主权利 (*Citizens' Personal, Democratic Rights*) **Subjective Element:** 因邻里纠纷(*due to a dispute*) **Objective Element:** **-- Harmful Action:** 刘某用尖刀将于某扎伤(*Liu stabbed Yu with a knife*) **-- Harmful Result:** 致使于某动脉破裂失血性休克死亡(*caused Yu femoral artery rupture shock death*)	**Subject Element:** 王某 (*Wang*) **Object Element:** 国家司法秩序 (*National Judicial Order*) **Subjective Element:** 被告人王某在明知其丈夫故意伤害他人的情况下 (*Wang known Liu injure others*) **Objective Element:** **-- Harmful Actions:** 仍帮助其逃跑 (*but still help him escape*) **-- Harmful Results:** 致使刘某逃避法律制裁 (*resulting in Liu to evade justice for 15 years*)

Figure 1: A single-defendant case (a) from CAIL2018 (Xiao et al., 2018) and a multi-defendant case (b) from our benchmark MUD with crime elements.

While in real-world scenarios a single case may involve multiple defendants, as shown in Fig. 1 (b) with two defendants, namely, *Liu* and *Wang*, whose criminal facts intertwine and overlap. Intuitively, addressing intricate cases with multiple defendants necessitates providing clear and compelling explanations for the criminal facts relevant to each defendant, ensuring the precision of charge predictions. Unfortunately, most of the existing datasets lack fine-grained annotations of criminal facts, consequently impairing the performance of current advanced methods.

As illustrated in Fig. 2 (a), several popular methods, *e.g.*, LegalBERT (Chalkidis et al., 2020), LawFormer (Xiao et al., 2021), and RoBERTa (Cui et al., 2021), show inferior performance on multiple-defendant cases (our new benchmark) compared to single-defendant cases (CAIL2018), with drops of 49%, 38%, and 32%, respectively. In this work, we construct a new benchmark with fine-grained annotations for multi-defendant legal cases, named MUD.

*corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2864–2878

Crime Elements	Definitions
犯罪客体(**Object Element**)	中华人民共和国刑法所保护的而为犯罪所侵害的人的社会生活利益(社会主义社会关系)。**(The interests of social life of the people protected by the criminal law of the People's Republic of China and infringed by the crime (socialist social relations).)**
犯罪客观方面(**Objective Element**)	犯罪活动的客观外在表现，包括危害行为、危害结果，行为与结果之间的因果关系。有些罪的构成还要求发生在特定的时间、地点或者使用特定的方法。**(The objective external manifestations of criminal activities, including harmful behaviors, harmful outcomes, and the causal relationship between behaviors and outcomes. The constitution of some crimes also requires the occurrence at a specific time, place, or use of specific methods.)**
犯罪主体(**Subject Element**)	实施犯罪行为,依法应当承担刑事责任的人,包括自然人、单位。**(Individuals who commit criminal acts and should bear criminal responsibility in accordance with the law, including natural persons and units.)**
犯罪主观方面(**Subjective Element**)	指行为人有罪过(包括故意和过失)。有些罪的构成还要求有特定的目的或动机。**(Refers to the perpetrator's guilt (including intent and negligence). The constitution of some crimes also requires a specific purpose or motive.)**

Table 1: Definition of four types of crime elements according to Criminal Law of the People's Republic of China. The translated version is indicated in bold font.

It focuses on multi-defendant criminal cases, comprising 2,865 cases and 7,128 defendant-charge pairs, spanning across 22 different charges. Moreover, each defendant in MUD is annotated with four types of crime elements. It is notable that, in the Chinese legal system, crime elements play a critical role in determining whether a particular action constitutes a crime or not (Cohen, 1982). These consist of the *Subject Element*, *Object Element*, *Objective Element*, and *Subjective Element*. Their precise definitions according to the Criminal Law of the People's Republic of China are shown in Table 1. By deeply understanding crime elements, legal professionals can more accurately determine criminal facts, ensuring fair trials. Similarly, in legal artificial intelligence (LegalAI) systems, the incorporation of crime elements is expected to enhance charge prediction accuracy, as well as improve model explainability and credibility. In Fig. 2 (b), our probing experiments confirm the efficacy of crime elements in boosting the performance of existing methods on MUD.

It is not trivial to annotate crime elements for each defendant. To ensure the quality, we adopted a three-stage annotation approach and engaged three legal experts. The experts' review of 500 randomly selected cases shows 99.3% agreement on annotated crime elements, confirming the high quality of our MUD benchmark.

Our new benchmark fills the gap in the absence of annotated crime elements, facilitating the creation of interpretable models. Based on MUD, we propose a new method named EJudge which jointly leverages crime elements and legal rules to infer charges. The extensive experiments show that MUD poses challenges to existing state-of-the-art models and verify the advancements of our method. Our contributions are as follows:

1) We propose a new multi-defendant charge prediction benchmark named MUD, in which four types of crime elements are annotated.

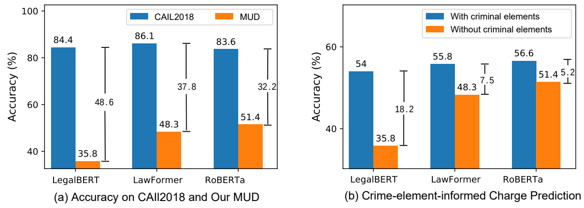

Figure 2: State-of-the-art models underperform on MUD compared with CAIL2018 (a). Their performance on our MELLE benchmark is significantly improved by incorporating the crime elements (b).

2) We design a crime-element-informed model named EJudge, which jointly leverages crime elements and legal rules to predict charges.

3) Extensive experiments verify the effectiveness of the proposed EJudge in leveraging crime elements with $+9.4\%$ F1 increase over existing methods for multi-defendant prediction.

2 Related Work

Legal Datasets. Recently, various datasets have been constructed for LegalAI, such as FLA (Luo et al., 2017), RACP (Jiang et al., 2018), Criminal (Hu et al., 2018), ECHR (Aletras et al., 2016), ECHR-Case (Chalkidis et al., 2019), ECHR-Crystal-Ball (Medvedeva et al., 2018), CAIL2018 (Xiao et al., 2018), QAjudge (Zhong et al., 2020) and FEDLEGAL (Zhang et al., 2023). However, these datasets mainly support coarse-grained charge prediction lacking detailed annotations.

To alleviate these problems, RACP (Jiang et al., 2018) and ACI (Paul et al., 2020) are constructed by randomly selecting 1,000 and 120 documents for sentence-level annotation, respectively. MNLM (Ge et al., 2021) provides fine-grained fact-article annotations for 1,189 legal cases, but there are only two charges. Yue et al. (2021b) constructs a dataset for charge prediction and court view generation that contains the crime circumstances.

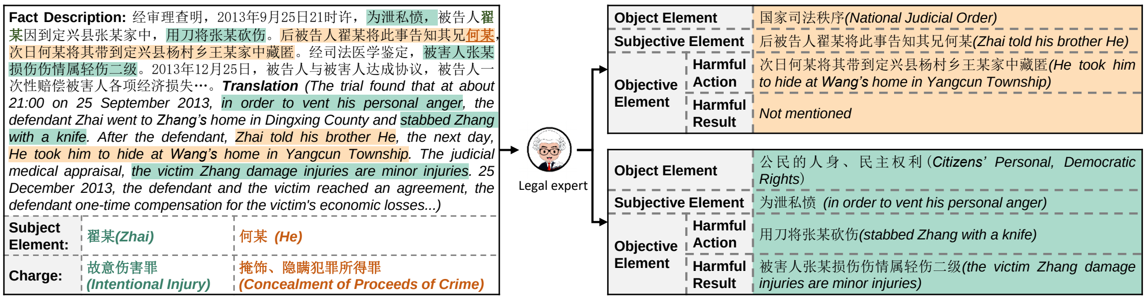

Figure 3: An example of an annotated case in MUD. For the given fact description, defendants, and corresponding charges (left), legal experts are required to select sentences mentioning constitutive elements (tables on the right).

SCE (An et al., 2022) provides sentence-level crime elements for 685 signal-defendant cases. Some works also delved into the practical scenarios of multi-defendant cases. MSA (Pan et al., 2019) is developed with the multi-scale attention model to predict the charge for each defendant. However, their used dataset only contains 100 legal cases. Later, MultiLJP (Lyu et al., 2023) is constructed for legal prediction containing a large-scale collection of multi-defendant cases. Following this line, we construct a new dataset with criminal elements for multi-defendant cases. In this work, we create a new benchmark consisting of 2,865 legal cases, with an average of 2.5 defendants per case and covering 22 different charges, and each defendant is annotated with crime elements.

Interpretable Methods. LegalAI has attracted attention in both research and practical applications, yielding notable achievements (Xiao et al., 2021; Feng et al., 2022). There has been a growing emphasis on the significance of interpretability in LegalAI, aiming to diminish the opacity of black-box models and improve the transparency of legal predictions (Jiang et al., 2018; Lyu et al., 2022; Zhao et al., 2022a; Li et al., 2022a; Luo et al., 2023; Barale et al., 2023). For example, Luo et al. (2017) show that manually designed ten elements such as *Violence* and *Death* are effective in distinguishing confusing charges. Jiang et al. (2018) and Zhong et al. (2020) verify the usefulness of contributory spans. Luo et al. (2023) make legal decisions by providing precedents and Legislations as inputs. Zhao et al. (2022b) design a multi-task learning method CPEE to explore the practical judicial process and analyzes comprehensive legal essential elements to make judgment predictions. NeurJudge (Yue et al., 2021a) separates the fact description into different circumstances and exploits them to make predictions. Recently, several works (Lyu et al., 2022; Zhao et al., 2022a; Li et al., 2022b; An et al., 2022) reveal the importance of crime elements for interpretable charge prediction. In line with this, we contribute a new benchmark annotated with crime elements and introduce a novel crime-element-informed method.

3 A New Benchmark: MUD

3.1 Data Collection

Our benchmark is sourced from China Judgment Online (CJO)[1], a Chinese government website that is widely used in LegalAI tasks (Xiao et al., 2018; Yao et al., 2022). We focus on multi-defendant cases. Specifically, we extract the fact description and defendant-charge pairs from the documents following Xiao et al. (2018). We discard fact descriptions with fewer than 50 characters or those involving only a single defendant. Then, charges with a frequency of less than 100 are filtered out. Through this process, we collect 2,856 documents containing 7,128 defendant-charge pairs covering 22 different charges for annotation.

3.2 Crime Elements Annotation

The identification of crime elements (as outlined in Table 1) is crucial in determining if a behavior constitutes a crime in the real-world conviction process (Cohen, 1982). This annotation process is conducted by senior Ph.D. students in law, who possess extensive legal knowledge and a comprehensive understanding of the four elements of crime.

Given a fact description and a defendant (*Subject Element*), annotators are required to label *Object Element* and select sentences mentioning *Objective Element* (*i.e.*, *Harmful Action*, *Harmful Result*) and *Subjective Element*. Fig. 3 shows an annotation example. The annotators are required to spend a minimum of 10 minutes on each fact and are compensated at a rate of $20 per hour based on the time required to complete the annotations.

[1] https://wenshu.court.gov.cn/

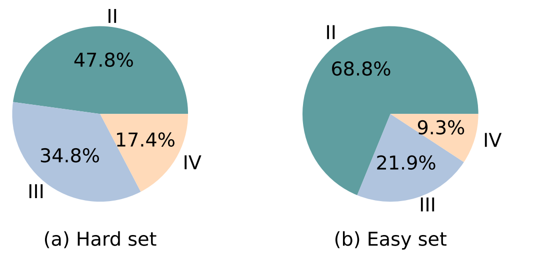

Figure 4: The MUD dataset was divided into two subsets: (a) *Hard*, where two or more defendants face different charges, and (b) *Easy*, where all defendants have the same charge. II, III, and IV denote cases with two, three, and four defendants, respectively.

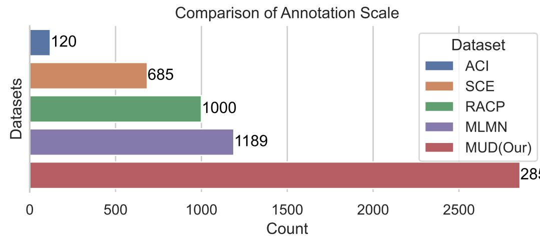

Figure 5: Comparison of the interpretable annotation scale of existing legal domain datasets (*e.g.*, RACP (Jiang et al., 2018), SCE (An et al., 2022), ACI (Paul et al., 2020), MLMN (Ge et al., 2021)), and our benchmark MUD.

Commonly used Legal Judgment Prediciton dataset CAIL-2018 (Xiao et al., 2018) relies on automatic extraction for annotation inevitably leading to some errors (as shown in Appendix A). In contrast, we design a three-stage annotation process. In the first stage, annotators are required to familiarize themselves with the annotation process by annotating a subset containing 500 cases that are randomly selected from MUD. Moving to the second stage, each case is annotated three times independently. We discard annotation results if the overlap ratio is less than 0.96. In the third stage, legal experts specifically focus on annotating cases discarded in the second stage, engaging in discussions to reach inter-annotator agreement.

3.3 Data Analysis

Dataset Statistics. There are 2,856 cases and 7,128 defendant-charge pairs covering 22 distinct charges in MUD. As shown in Fig. 4, we divide MUD into two subsets: the *Easy* set, where each case involves all defendants accused of the same charge; and the *Hard* set, where at least two defendants face different charges, posing a greater challenge for charge prediction.

Dataset Quality. To evaluate the dataset quality, we randomly sample 500 cases labeled three times independently from MUD. The legal experts' review shows 99.3% agreement on annotated crime elements, demonstrating that the MUD is a high-quality manually annotated benchmark.

Annotation Scale. The annotation process for datasets in the legal domain is complex and rigorous, and annotators are required to possess a strong legal background. As far as we know, our new benchmark provides the largest fine-grained annotation scale for multi-defendant charge prediction. Some legal datasets also provide fine-grained annotations beside the charge labels. Fig. 5 shows the annotations scale of MUD and existing fine-grained annotated datasets in the legal domain. SCE (An et al., 2022) is annotated with sentence-level criminal elements for 685 cases, but they only consider the signal-defendant cases from CAIL. RACP (Jiang et al., 2018) and ACI (Paul et al., 2020) randomly select 1,000 and 120 documents for sentence-level annotation, respectively. MNLM (Ge et al., 2021) provides fine-grained fact-article annotations for 1,189 legal cases covering two different charges. Our benchmark MUD provides crime element annotations for 2,856 cases, which is much larger than the existing datasets.

4 Crime-Element-Informed Method

4.1 Task Definition

Given a multiple-defendant case, its fact description is denoted as f and the involved defendants are denoted as $\mathcal{D} = \{d_1, d_2, ..., d_l\}$, where l is the number of defendants. The task is to predict the charge for each defendant. To enhance interpretability, legal knowledge is incorporated into the prediction process. In this work, we leverage category information for the legal system, denoted as $\mathcal{C} = \{C_1, C_2, \cdots, C_m\}$, where C_i represents a crime category encompassing n_i charges, *i.e.*, $C_i = \{c_{i,1}, c_{i,2}, \cdots, c_{i,n_i}\}$. Additionally, we use the legal rules $\mathcal{R} = \{r_{1,1}, r_{1,2}, .., r_{m,n_m}\}$ defined by law, where $r_{i,j}$ is the legal rule of charge $c_{i,j}$.

4.2 EJudge

Overview. Fig. 6 shows the overall architecture of EJudge. The basic idea is to deduce the charges against each defendant by analyzing the elements of the crime in conjunction with relevant legal rules. The *Element Selector* extracts crime elements for each defendant from the fact description. Subse-

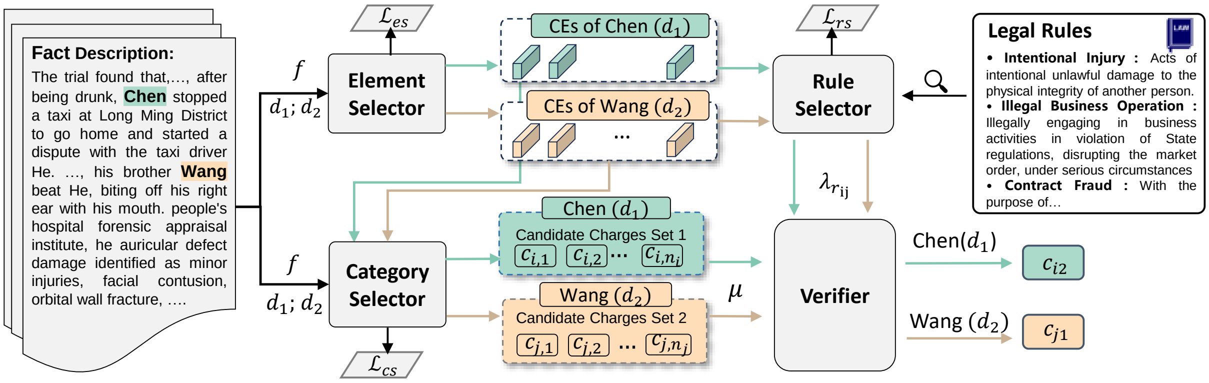

Figure 6: Overall architecture of EJudge. EJudge consists of four components: the *Element Selector*, the *Category Selector*, the *Rule Selector*, and the *Verifier*. CEs denote the crime elements.

quently, the *Category Selector* predicts charge categories, and the *Rule Selector* improves the differentiation of confusing charges within each category using legal rules. The *Verifier* integrates predicted charge categories and legal rules to infer charges. We detail the four modules below.

Element Selector. The conviction process is rigorous and requires consideration of the crime elements in the facts. This module aims to extract the sentences mentioning the crime elements for a given defendant. First, for each defendant d_i, we generate the representation of defendant-aware fact description f by passing them into a pre-trained encoder (*e.g.*, RoBERTa (Cui et al., 2021)):

$$\mathbf{H}_{d_i} = \text{Encoder}(\texttt{[CLS]}\ d_i\ \texttt{[SEP]}\ f\ \texttt{[SEP]}), \quad (1)$$

where `[CLS]` and `[SEP]` are the special tokens, and $\mathbf{H}_{d_i}$ is the output embeddings for all input tokens. We use the NLTK tool[2] to split the fact description into sentences, *i.e.*, $f = \{s_1, s_2, s_3, \cdots\}$, and generate the sentence embeddings using an average-pooling layer:

$$\mathbf{h}_{s_i} = \text{avg\_pool}(\mathbf{h}_{w_{i,1}}, \mathbf{h}_{w_{i,2}}, ..., \mathbf{h}_{w_{i,j}}, ...), \quad (2)$$

where $w_{i,j}$ is the j-th word in sentence s_i, $\mathbf{h}_{w_{i,j}}$ is its word embedding in $\mathbf{H}_{d_i}$, and $\mathbf{h}_{s_i}$ is the sentence embedding of s_i. Then, we apply a linear classifier on $\mathbf{h}_{s_i}$ followed by a softmax function to predict the element probabilities $\hat{\mathbf{p}} \in \mathbb{R}^{K_e}$ for four types of elements, where K_e is the number of element types, *i.e.*, 4. We train the module by the element classification loss, which can be formulated as:

$$\mathcal{L}_{es} = \mathbb{E}[-\sum_{k_e=1}^{K_e} \mathbf{p}(k_e|h_{s_i}) \log(\hat{\mathbf{p}}(k_e|h_{s_i}))], \quad (3)$$

where $\mathbb{E}$ denotes the average expectation, and $\mathbf{p}(k_e|s_i)$ represents the ground-truth probability of crime elements based on the sentence s_i. For the ground-truth crime element class k_e, the $\mathbf{p}(k_e|s_i)$ equals to 1 otherwise 0.

Category Selector. In the legal domain, charges are divided into different categories depending on the *Object Element*. Appendix C shows several examples of charge categories. Generally, given the fact, it's easy to identify the crime categorize, such as distinguishing between *Public Social Security* and *Market Economic Order*). However, the difficulty arises when trying to differentiate between confusing charges within the same category, such as *Intentional Homicide* and *Involuntary Manslaughter*. Inspired by this observation, we first predict the charge category for each defendant. Specifically, for each defendant d_i, we obtain the embedding sequence of the fact description as defined in Eq. (1), and employ an average-pooling layer to get the defendant-aware fact description, denoted as $\mathbf{h}_f$. Then, we use the same encoder to encode the extracted crime elements for d_i, and leverage an average pooling layer over the output sequence embeddings to get the context representation of crime elements, denoted as $\mathbf{h}_e$. We concatenate $\mathbf{h}_f$ and $\mathbf{h}_e$ and pass them through a linear layer as the category feature μ_{d_i}. We train the module by the category classification loss as:

$$\mathcal{L}_{cs} = \mathbb{E}[-\sum_{k_c=1}^{K_c} \mathbf{p}(k_c|\mu_{d_i}) \log(\hat{\mathbf{p}}(k_c|\mu_{d_i}))], \quad (4)$$

where $\mathbb{E}$ denotes the average expectation, and K_c is the number of charge categories. $\hat{\mathbf{p}}(k_c|\mu_{d_i}) \in \mathbb{R}^{K_c}$ denote the predicted category probabilities, and $\mathbf{p}(k_c|\mu_{d_i})$ represent the ground-truth probability of charge categories based on the category feature μ_{d_i}, which equals to 1 otherwise 0.

Rule Selector. In our method, convictions are based on aligning crime elements with the relevant

[2] https://www.nltk.org/

legal rules. This module is designed to calculate matching scores between legal rules in the selected categories and extracted crime elements, identifying the most probable legal rules for charging. The legal rule of each charge is clearly defined, and several examples are shown in Appendix D. In this module, we use the pre-trained encoder (*e.g.*, RoBERTa (Cui et al., 2021)) to separately encode the word sequence of the legal rule $r_{i,j}$, and sentences containing crime elements of defendant d_i. Then we obtain the hidden vector sequence of the legal rule $\mathbf{H}_r = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{l_r}\}$ and crime elements $\mathbf{H}_e = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{l_e}\}$, where l_r and l_e represent the sequence length. We apply an average-pooling layer to $\mathbf{H}_r$ and $\mathbf{H}_e$ to get the embedding of the legal rule and crime elements, which are denoted as $\mathbf{h}_{r_{i,j}}$, and $\mathbf{h}_e$, respectively. Finally, we use the cosine similarity function to measure the matching score between them:

$$\lambda_{r_{i,j}} = \frac{\mathbf{h}_e \cdot \mathbf{h}_{r_{i,j}}}{|\mathbf{h}_e| \cdot |\mathbf{h}_{r_{i,j}}|}, \tag{5}$$

where $\lambda_{r_{i,j}}$ represents the matching score. We train the *Rule Selector* by optimizing contrastive loss. Specifically, given the legal rule $r_{i,j}$ of charge $c_{i,j}$, we sample sentences mentioning crime elements of the charge $c_{i,j}$ as s^+. We sample sentences mentioning crime elements of charge $c_{i,t} (t \neq j)$, which we denote as s^-. With above steps , we construct positive pairs $(r_{i,j}, s^+)$ and negative pairs $(r_{i,j}, s^-)$. The contrastive loss is defined as:

$$\mathcal{L}_{rs} = -log \frac{e^{sim(\mathbf{h}_{r_{i,j}}, \mathbf{h}_{s^+})/\tau}}{\sum_{j=1}^{N} (e^{sim(\mathbf{h}_{r_{i,j}}, \mathbf{h}_{s^+})/\tau} + e^{sim(\mathbf{h}_{r_{i,j}}, \mathbf{h}_{s^-})/\tau})}, \tag{6}$$

where N, τ, and sim represent the mini-batch size, temperature hyperparameter, and cosine similarity function, respectively.

Verifier. This module aims to aggregate the scores generated by the *Category Selector* and *Rule Selector* to make a final decision. Specifically, we select categories with the top-η highest logits generated by the *Category Selector*, where the selected categories set is denoted as $\mathcal{C}'_\eta = \{C'_1, C'_2, ..., C'_\eta\}$. We choose the charge for which the corresponding legal rule has the highest probability $p_{r_{i,j}}$ as the final charge prediction $\hat{c}_{i,j}$:

$$q(k_c|\mu_{d_i}) = softmax(\alpha\mu_{d_i}), \tag{7}$$

$$q(k_r|\lambda_{r_{i,j}}) = softmax(\beta\lambda_{r_{i,j}}), \tag{8}$$

$$\arg\max_{i,j} \{p_{r_{i,j}} | p_{r_{i,j}} = q(k_c|\mu_{d_i}) * q(k_r|\lambda_{r_{i,j}}), k_c \in \mathcal{C}'_\eta\}, \tag{9}$$

Dataset	MUD			CAIL-2018
	Easy	Hard	All	
#Train	1,184	555	1,739	101,275
#Dev	387	169	556	-
#Test	386	175	561	26,661

Table 2: Statistics of MUD and CAIL, where "#" denotes the number of data in the set.

where $\lambda_{r_{i,j}}$ represents the similarity score which is generated by the *Rule Selector* (Eq. 5), α and β are the temperature hyperparameters.

5 Experiment

5.1 Experiment Setting

Dataset and Metrics. We conduct experiments for multi-defendant charge prediction on our MUD, which is randomly split into the training set, development set, and test set, following a ratio of 3 : 1 : 1. We also conduct experiments on the commonly used dataset CAIL (Xiao et al., 2018) with single-defendant cases to verify the effectiveness of EJudge. The details of used datasets are shown in Table 2.

Each case in MUD contains multiple defendants. Following Lyu et al. (2022), we adopt Accuracy (Acc), Macro Precision (MaP), Macro Recall (MaR), and Macro F1 (MaF) to evaluate the model's ability to predict charges for defendants. In addition, we use Accuracy (Acc*) to measure the model's ability to predict the charge for cases, *i.e.*, whether correctly assign the charge for all defendants in a case.

Baseline Models. To verify the effectiveness of our model Ejudge, we compare EJudge with the following methods which are summarized in the three groups: **Single-Defendant Methods** including **DPAM** (Wang et al., 2018), which incorporate law articles to help charge prediction; **CECP** (Zhao et al., 2022a), **DCSCP** (Li et al., 2022a) and **GEEN** (Lyu et al., 2022), which predict charges by extracted crime elements; **HMN** (Wang et al., 2019), which formulates charge prediction as a hierarchical multi-label classification problem; **NeurJudge** (Yue et al., 2021a), which splits the facts into several parts to predict charges; **CTM** (Liu et al., 2022), which takes case triples as input to predict charges. **Multi-Defendant Methods** including **MSA** (Pan et al., 2019), which predict charge by using a multi-scale attention model; **Pre-trained Language Models** including **RoBERTa**,

	Models	Hard(%)					Easy(%)					All(%)				
		Acc	MaP	MaR	MaF	Acc*	Acc	MaP	MaR	MaF	Acc*	Acc	MaP	MaR	MaF	Acc*
w/o E	MSA	63.4	51.1	50.6	49.1	36.6	78.2	78.5	78.3	78.2	77.9	75.4	75.1	74.6	75.0	62.1
	CECP	60.1	52.2	51.3	49.8	35.9	80.0	80.7	80.5	80.4	80.9	76.1	76.1	76.0	76.0	65.0
	DCSCP	61.4	51.1	50.9	49.9	36.5	78.9	80.5	80.6	80.3	80.1	76.2	76.1	76.1	76.0	64.0
	LeaglBERT	62.4	52.0	48.5	48.4	35.8	80.1	81.0	80.2	80.2	79.8	76.8	76.5	76.7	76.5	64.3
	RoBERTa	**74.0**	**66.6**	**64.8**	**64.6**	**51.4**	82.0	82.6	82.0	82.0	81.2	79.6	**80.6**	80.4	80.4	68.2
	LawFormer	72.4	63.4	58.9	58.9	48.3	**84.5**	**84.9**	**84.9**	**83.6**	**83.8**	**80.2**	80.5	**81.4**	**81.4**	**68.5**
w/ E	DPAM*	73.5	67.9	61.3	60.8	50.4	81.9	81.1	81.0	82.2	80.1	77.8	76.9	77.2	76.1	71.3
	HMN*	73.3	67.4	64.1	66.2	50.8	82.1	83.3	83.5	83.6	81.1	79.3	80.2	80.2	80.6	70.1
	CTM*	74.5	66.2	64.0	60.0	51.1	83.9	82.5	83.1	82.7	81.3	81.8	81.3	81.3	81.6	69.7
	GEEN*	73.0	65.1	64.4	59.2	50.6	84.6	83.2	83.2	83.5	81.9	82.1	81.9	82.5	82.6	70.7
	NeurJudge*	73.8	67.8	65.6	67.9	52.1	83.1	83.2	92.9	83.4	80.5	79.0	81.5	81.5	80.9	71.1
	LegalBERT*	72.4	60.4	60.9	61.8	51.0	79.8	80.1	79.6	80.4	80.2	76.3	74.2	75.6	75.6	70.2
	RoBERTa*	**74.8**	67.4	65.8	66.0	**54.8**	82.6	83.2	83.0	82.4	81.8	81.0	81.8	80.8	81.0	71.0
	LawFormer*	74.5	65.3	62.2	62.4	53.6	84.5	85.0	84.8	83.5	**82.3**	81.3	82.2	81.7	82.0	69.0
	EJudge*	74.5	**74.8**	**74.8**	**74.0**	54.0	**85.8**	**86.4**	**86.3**	**86.1**	82.0	**82.6**	**83.0**	**82.9**	**82.9**	**71.3**
Oracle	DPAM$^+$	74.5	72.6	69.0	67.3	51.8	83.3	83.6	83.1	83.4	81.6	80.2	80.3	80.3	80.1	72.8
	HMN$^+$	74.7	70.4	67.1	68.0	52.1	83.5	84.0	84.2	84.5	82.9	79.4	80.9	80.5	81.0	74.3
	CTM$^+$	75.2	71.2	65.2	66.7	54.8	83.3	83.6	83.1	83.4	82.6	80.3	81.0	81.0	81.2	74.8
	GEEN$^+$	75.1	72.7	65.8	67.1	55.2	84.8	83.9	83.5	83.6	83.6	81.5	81.3	81.3	81.3	75.1
	NeurJudge$^+$	76.5	73.4	68.0	69.4	57.3	84.2	84.5	85.0	85.6	83.8	80.9	82.3	81.6	81.6	72.4
	LegalBERT$^+$	75.1	78.4	71.3	71.0	54.0	80.2	79.9	79.3	79.3	79.0	77.3	76.5	77.3	77.4	71.8
	RoBERTa$^+$	77.6	73.8	69.4	69.8	56.6	82.4	83.6	83.0	82.6	82.4	80.6	80.3	80.5	80.5	76.4
	LawFormer$^+$	76.5	74.8	66.4	68.2	55.8	85.5	85.8	85.4	85.0	84.3	83.3	83.2	82.9	82.9	76.1
	EJudge$^+$	**78.0**	**81.0**	**80.9**	**78.4**	**59.8**	**87.3**	**87.9**	**87.9**	**87.6**	**84.5**	**84.0**	**84.4**	**84.7**	**84.4**	**77.3**

Table 3: Overall performance of multi-defendant charge prediction on MUD. The best results under different settings are marked in **bold**. *w/o* and *w* E denote whether we explicitly extract crime elements in facts for prediction.

Models	CECP	DCSCP	DPAM	HMN	GEEN	NeuralJudge	LegalBERT	RoBERTa	LawFormer	**EJudge**
Acc	0.8651	0.8599	0.8462	0.8298	0.8433	0.8565	0.8432	0.8360	0.8679	**0.8688**
MaP	0.8511	0.8323	0.8407	0.8323	0.8587	0.8634	0.8502	0.8412	**0.8702**	0.8691
MaR	0.8632	0.8677	0.8512	0.8434	0.8489	0.8413	0.8356	0.8322	0.8544	**0.8634**
MaF	0.8533	0.8572	0.8415	0.8399	0.8519	0.8511	0.8444	0.8398	0.8633	**0.8652**

Table 4: Overall performance of single-defendant charge prediction on CAIL. The best results are marked in **bold**.

which is pre-trained language model of Chinese version (Cui et al., 2021); **LegalBERT** (Zhong et al., 2019) and **LawFormer** (Xiao et al., 2021), which are pre-trained language models in the legal domain. Moreover, we also explore the performance of large language models (LLMs) on our MUD, including **GPT-4.0** (OpenAI, 2023), **GPT-3.5** (Ouyang et al., 2022), and **GLM-130B** (Zeng et al., 2023).

Implementation Details. We use the released source codes to implement baseline models (*i.e.*, DPAM, HMN, NeurJudge, CTM, MSA, GEEN, CECP, DCSCP, HRN). For EJudge, we set the dropout rate, learning rate, batch size, warmup steps, and max length of fact as 0.1, 1×10^{-5}, 12, 800, and 500, respectively. For the *Rule Selector*, we sample four negative samples and one positive sample for each instance. We search τ in Eq. (6), top-η of *Verifier*, α in Eq. (7), and β in Eq. (8) with grid searching, where $\tau \in \{0.01, 0.05, 0.1\}$, top-$\eta \in \{1, 3, 5\}$, $\alpha \in \{0.1, 0.3, 0.5\}$, and $\beta \in \{0.1, 0.3, 0.5\}$. The implementation is based on Pytorch and trained on a Tesla V100 GPU with AdamW (Loshchilov and Hutter, 2019) optimizer for 20 epochs. We choose the checkpoints with the best average performance on the development set and report performance on the test set. In terms of crime elements, the experiments are implemented under three settings: **(1) Without elements** (*w/o* E): Only the fact description is used for charge prediction. **(2) With extracted elements** (*w/* E, marked with "$*$"): The fact description and extracted crime elements are used for charge prediction. **(3) With annotated elements** (Oracle, marked with "$+$"): The fact description and annotated crime elements are used for charge prediction.

5.2 Main Results

Table 3 shows the overall performance on MUD. It is observed that EJudge outperforms all baselines by a large margin. For example, in the *Hard* set of MUD, our EJudge outperforms the prior SOTA method without elements (*i.e.*, RoBERTa) by improving MaF by 9.4% and 13.8% using extracted and annotated elements, respectively. Table 4 reports the performance of single-defendant charge prediction on CAIL (Xiao et al., 2018). It is observed that the element-aware methods, *i.e.*, CECP, DCSCP, GEEN, and EJudge, surpass other meth-

Dateset	Model	Acc	Map	MaR	MaF	Acc*
Hard	**EJudge***	**74.5**	**74.8**	**74.8**	**74.0**	54.0
	-ES	70.4	67.4	65.8	66.0	50.8
	-CS	74.1	70.4	70.2	68.9	**54.2**
	-RS	72.3	69.2	69.5	69.6	53.8
	-V	72.1	69.8	71.8	69.6	52.0
Easy	**EJudge***	**85.8**	**86.4**	**86.3**	**86.1**	**82.0**
	-ES	82.6	83.2	83.0	82.4	81.8
	-CS	85.6	83.4	83.5	84.1	80.9
	-RS	82.4	81.9	82.0	81.9	79.8
	-V	85.1	85.6	85.5	85.3	81.1

Table 5: Ablation study on the test set of MUD.

ods by 1.7% in terms of average MaR, showing the importance of crime elements. Moreover, EJudge achieves the best performance, indicating the superiority of our method in leveraging crime elements. Furthermore, we explore the performance of LLMs under zero- and few-shot settings on the full test dataset of MUD. The best ACC* achieved by LLMs is 59.73%, worse than all baseline models trained with labeled cases. This indicates LLMs' limitations in dealing with professional and intricate legal scenarios. Please refer to Appendix E for details about experiment settings and results analysis.

5.3 In-Depth Analysis

Ablation Study. We conduct an ablation study to illustrate the effectiveness of each component of EJudge*. Table 5 shows that removing the *Element Selector* (-ES), *Category Selector* (-CS), *Rule Selector* (-RS), or *Verifier* (-V) leads to performance drops, indicating each component is useful for multi-defendant charge prediction.

Element Selector. In Table 3, the performance of EJudge$^+$ (which utilizes annotated crime elements) surpasses that of EJudge* (which relies on extracted elements), by an average margin of 3.68%, showing that enhancing the quality of extracted crime elements can benefit charge prediction. To investigate the quality of extracted elements, we fine-tune LegalBERT (Chalkidis et al., 2020), LawFormer (Xiao et al., 2021), and RoBERTa (Cui et al., 2021) in the *Element Selector* module, and report exact match scores at both sentence- and token-level in Table 6. The averaged exact match scores are 74.1% and 75.6% on sentence- and token-level respectively, indicating scope for improvement.

Category Selector and Rule Selector. Considering the charges of *Fraud*, *Contract Fraud*, and *Extortion*, we investigate the model's ability to distinguish confusing charges. Fig. 7 shows that when removing the *Category Selector* (a) from EJudge* (*w/o* CS), it is hard to distinguish confusing charges (*Fraud* and *Contract Fraud*) in different categories. Removing the *Rule Selector* (b) makes it difficult to differentiate between *Fraud* and *Extortion*, which are in the same category. These observations verify the effectiveness of the *Rule Selector* and *Category Selector* for accurate charge prediction, by leveraging interpretable crime elements and legal rules.

Models	Hard(%)		Easy(%)	
	Sent	Token	Sent	Token
LegalBERT	71.4	71.9	73.5	75.6
LawFormer	73.5	74.9	**76.4**	77.1
RoBERTa	**74.2**	**75.7**	75.8	**77.3**

Table 6: Results of crime element extraction. *Sent* denotes sentence-level exact match, and *Token* denotes token-level exact match.

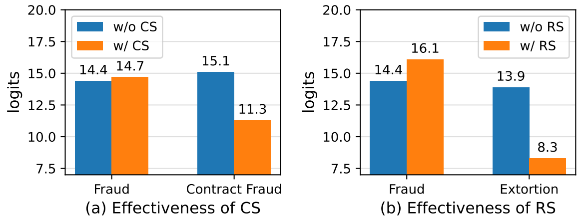

Figure 7: The *Category Selector* (a) benefits for distinguishing *Fraud* and *Contract Fraud* that are in the different charge categories. The *Rule Selector* (b) benefits for distinguishing *Fraud* and *Extortion* that are in the same charge categories. CS and RS denote the *Category Selector*, and *Rule Selector*, respectively.

Case Study and Interpretability Analysis. Fig. 8 presents a case with three defendants *Zhu*, *Jiang*, and *Wang* whose criminal facts intertwine and overlap. The existing methods, such as LegalBERT (Chalkidis et al., 2020), and EJudge*-ES can correctly predict the charges relevant to the whole case but fail to accurately assign the charges for each defendant. Our model EJudge* correctly predicts the charge for each defendant. Notably, our EJudge* method provides the extracted crime elements and matched legal rules, enhancing both prediction accuracy and model interpretability.

6 Conclusion

In this study, we introduce a new charge prediction benchmark called MUD that comprises multi-defendant legal cases. We annotate the crime elements for each defendant, which benefits interpretable model development. Moreover, we

事实描述： …，朱某在淄博市张店区中段路口因争抢出租车与邢某发生争执，朱某联系江某，江某伙同王某持木棍、匕首赶到现场，与邢某纠集的张某、翟某等人发生殴斗，在殴斗过程中，王某持匕首将翟某腹部、张某腿部捅伤。经鉴定，翟某伤情构成重伤二级、张某伤情构成轻伤二级。… ***#Translation*** **(Fact Description:** …, ***Zhu*** *had a dispute with Xing for fighting for a taxi. Zhu contacted* ***Jiang****. Jiang and* ***Wang*** *arrived at the scene with sticks and daggers, and had a fight with Zhang and Zhai gathered by Xing. During the fight, Wang stabbed Zhai in the abdomen and Zhang in the leg with a dagger. After identification, Zhai and Zhang's injury constitutes a serious injury of grade II...)*

Models	Charge Prediction			Crime Elements
	D-1: 王某(Wang)	D-2: 江某(Jiang)	D-3:朱某(Zhu)	
LeaglBERT	Affray ✗	Intentional Injury ✗	Affray ✓	NO
EJudge*-ES	Intentional Injury ✓	Intentional Injury ✗	Affray ✓	NO
EJudge*(**Our**)	Intentional Injury ✓	Affray ✓	Affray ✓	**YES**

Crime Elements Extracted by EJudge*				
		D-1: 王某(Wang)	D-2: 江某(Wang)	D-3:朱某(Zhu)
SE		*Not Mentioned*	*Not Mentioned*	朱某因争抢出租车与邢某发生争执*(Zhu had a dispute with Xing for fighting for a taxi)*
OE	**HA**	王某持匕首将翟某腹部、张某腿部捅伤 *(Wang stabbed Zhai in the abdomen and Zhang in the leg with a dagger.)*	江某伙同王某持木棍、匕首赶到现场 *(Zhu Jiang and Wang arrived at the scene with sticks and daggers)*	朱某联系江某，江某伙同王某持木棍、匕首赶到现场*(Zhu contacted Jiang. Jiang and Wang arrived at the scene with sticks and daggers)*
	HR	翟某构成重伤二级、张某构成轻伤二级*(Zhai and Zhang's injury constitutes a serious injury of grade II...)*	与邢某纠集的张某，翟某等人发生殴斗 *(and had a fight with Zhang and Zhai gathered by Xing)*	与邢某纠集的张某、翟某等人发生殴斗 *(had a fight with Zhang and Zhai gathered by Xing)*

Figure 8: An example of a charge prediction. D, SE, and OE denote the Defendant, *Subjective Element*, and *Objective Element*, respectively. *Objective Element* contains *Harmful Action* (HA) and *Harmful Results* (HR).

propose a crime-element-informed model named EJudge, which outperforms existing methods for multi-defendant charge prediction. In the future, we will work on more accurate crime element extraction for interpretable charge prediction.

Limitations

In this work, we aim to promote the development of LegalAI, providing a new benchmark with annotated crime elements to the community. The limitation of this work is that the proposed EJudge represents an initial exploration into incorporating crime elements for charge prediction. In EJudge, we integrate crime elements by directly concatenating implicit representations of extracted elements with fact descriptions. Although this straightforward method shows the advantage of crime elements, the potential to take full advantage of these constitutive elements is still under-explored.

Ethics Statement

Each case included in the MUD benchmark has been obtained from the Chinese government website, with sensitive information appropriately anonymized to protect privacy. During the document selection stage, we filter out any segments that might contain personal information, such as name, gender, age, address, and more. For the annotation task, we initially annotated a subset of cases ourselves, and then we established annotator wages based on local standards to ensure fair compensation. It is important to note that while our work aims to alleviate the workload of legal professionals, our LegalAI model, like any other, may occasionally make mistakes. Therefore, we emphasize that our model should only serve as an additional auxiliary tool in the legal field. The ultimate decision-making should always depend on legal professionals.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant No.62302287. The research reported in this article was supported by the Shanghai Committee of Science and Technology, China (Grant No.23ZR1423500). We thank the anonymous reviewers for their insightful comments.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Zhenwei An, Quzhe Huang, Cong Jiang, Yansong Feng, and Dongyan Zhao. 2022. Do charge prediction models learn legal theory? In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Claire Barale, Michael Rovatsos, and Nehal Bhuta. 2023. Automated refugee case analysis: A NLP pipeline for supporting legal practitioners. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2992–3005.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. Senticnet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 3829–3839. European Language Resources Association.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4317–4323. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics.

Jerome Alan Cohen. 1982. The criminal procedure law of the people's republic of china. *The Journal of Criminal Law and Criminology (1973-)*, 73(1):171–203.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664, Dublin, Ireland. Association for Computational Linguistics.

Jidong Ge, Yunyun Huang, Xiaoyu Shen, Chuanyi Li, and Wei Hu. 2021. Learning fine-grained fact-article correspondence in legal cases. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3694–3706.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xin Jiang, Hai Ye, Zhunchen Luo, WenHan Chao, and Wenjia Ma. 2018. Interpretable rationale augmented charge prediction system. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 146–151, Santa Fe, New Mexico. Association for Computational Linguistics.

Lin Li, Lingyun Zhao, Peiran Nai, and Xiaohui Tao. 2022a. Charge prediction modeling with interpretation enhancement driven by double-layer criminal system. *World Wide Web*, 25(1):381–400.

Lin Li, Lingyun Zhao, Peiran Nai, and Xiaohui Tao. 2022b. Charge prediction modeling with interpretation enhancement driven by double-layer criminal system. *World Wide Web*, 25(1):381–400.

Dugang Liu, Weihao Du, Lei Li, Weike Pan, and Zhong Ming. 2022. Augmenting legal judgment prediction with contrastive case relations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2658–2667, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark. Association for Computational Linguistics.

Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Prototype-based interpretability for legal citation prediction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4883–4898.

Yougang Lyu, Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang, and Zhaochun Ren. 2023. Multi-defendant legal judgment prediction via hierarchical reasoning. pages 2198–2209. Association for Computational Linguistics.

Yougang Lyu, Zihan Wang, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Xiaozhong Liu, Yujun Li, Hongsong Li, and Hongye Song. 2022. Improving legal judgment prediction through reinforced criminal element extraction. *Information Processing & Management*, 59(1):102780.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Trans. Affect. Comput.*

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the european court of human rights: Looking into the crystal ball. In *Proceedings of the conference on empirical legal studies*, page 24.

Inderjeet Nair and Natwar Modani. 2023. Exploiting language characteristics for legal domain-specific language model pretraining. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2516–2526, Dubrovnik, Croatia. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sicheng Pan, Tun Lu, Ning Gu, Huajuan Zhang, and Chunlin Xu. 2019. Charge prediction for multi-defendant cases with multi-scale attention. In *Computer Supported Cooperative Work and Social Computing: 14th CCF Conference, ChineseCSCW 2019,* *Kunming, China, August 16–18, 2019, Revised Selected Papers 14*, pages 766–777. Springer.

Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. 2020. Automatic charge identification from facts: A few sentence-level charge annotations is all you need. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1011–1022, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7337–7348.

Pengfei Wang, Yu Fan, Shuzi Niu, Ze Yang, Yongfeng Zhang, and Jiafeng Guo. 2019. Hierarchical matching network for crime classification. In *proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 325–334.

Pengfei Wang, Ze Yang, Shuzi Niu, Yongfeng Zhang, Lei Zhang, and ShaoZhang Niu. 2018. Modeling dynamic pairwise attention for crime classification over legal articles. In *the 41st international ACM SIGIR conference on research & development in information retrieval*, pages 485–494.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *arXiv preprint arXiv:2105.03887*.

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*.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale Chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 183–201, Dublin, Ireland. Association for Computational Linguistics.

Wei Jie Yeo, Ranjan Satapathy, Rick Siow Mong Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? *CoRR*, abs/2402.11863.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021a. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 973–982.

Linan Yue, Qi Liu, Han Wu, Yanqing An, Li Wang, Senchao Yuan, and Dayong Wu. 2021b. Circumstances enhanced criminal court view generation. In *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1855–1859.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu, and Zenglin Xu. 2023. FEDLEGAL: The first real-world federated learning benchmark for legal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3492–3507.

Jie Zhao, Ziyu Guan, Cai Xu, Wei Zhao, and Enze Chen. 2022a. Charge prediction by constitutive elements matching of crimes. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 4517–4523.

Lili Zhao, Linan Yue, Yanqing An, Yuren Zhang, Jun Yu, Qi Liu, and Enhong Chen. 2022b. CPEE: civil case judgment prediction centering on the trial mode of essential elements. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management(CIKM)*, pages 2691–2700.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively Questioning and Answering for Interpretable Legal Judgment Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1250–1257.

Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Open chinese language pre-trained model zoo. Technical report.

Mislabeling

"**fact**": "吉林省蛟河市人民检察院指控：（一）、2012年12月份至2013年3月间，被告人桑某某以自家无烧柴为由，多次携带刀锯在某某镇某某村某某屯某蛟河市某某林场某某区122林班、南山蛟河市某某林场某某区132林班国有林内，盗窃国有木材柞树、杨树、白桦树、胡桃楸树、黄菠椤树，核原木材积10.338立方米，价值人民币4，135.00元。…追究其刑事责任。"
"**accusation**": ["故意伤害","盗窃","非法采伐、毁坏国家重点保护植物"]

Non-existent Label

"**fact**":"荣成市人民检察院指控，2016年春至10月8日期间，被告人宋某在荣成市斥山街道办事处夏家泊村等地附近的地里，利用细犬等狩猎野生动物，共猎捕24只野兔、1条虎斑颈槽蛇、1条三线蛇、1条土脚蛇，所猎捕野生动物共计价值人民币2400元。"
"**accusation**": ["非法狩猎"]

Figure 9: Two error types in LegalAI dataset CAIL-2018 (Xiao et al., 2018). For Mislabeling error type, the automatic method incorrectly asign the crime of intentional injury to the case，when in fact the case did not involve intentional injury. For Non-existent Label error type, as far as we know. The Illegal Hunting is undefined in the Criminal Law of China.

A Errors in CAIL-2018

Commonly used dataset cail-2018 (Xiao et al., 2018) in LegalAI task relies on automatic extraction for annotation, which inevitably leads to some errors. As shown in Fig. 9, we list out some error types, i.e. Mislabeling and Non-existent label. Mislabeling refers to labeling the case with the wrong charge. Non-existent label means labeling the case with the charge that undefined in the Criminal Law of China.

B Existing Datasets

In LegalAI field, there are several wildly used datasets. To compare with our MUD, We summarize the existing datasets in Table 7. In the early stage, Most of the works, such as Xiao et al. (2018), and QAjudge (Zhong et al., 2020), focus on single-defendant charge prediction. Recently, Lyu et al. (2023) construct a new legal judgment prediction dataset, where each criminal case contains an average of 3.4 defendants. However, it mainly supports black-box model development. In our study, we propose a new benchmark with high-quality crime element annotation, which can support interpretable model development.

C Category of Charges

Charges are arranged into different categories according to the Criminal Law of China, as shown in Fig. 12.

D Rule of Charges

The rule of charges expresses the conviction process, and the specific crime elements have corresponding formal terms in the rule. Table 8 shows their definitions according to the Criminal Law of China.

E Charge Prediction via Large Language Models

Large Language Models (LLMs) have shown remarkable performance in many domain-specific tasks, such as the sentiment analysis (Yeo et al., 2024; Mao et al., 2023; Cambria et al., 2022), and law domain (Shui et al., 2023). In this section, we conduct zero and few-shot experiments to evaluate LLMs on our benchmark MUD. We hope that these results can supplement previous research on the multi-defendant charge prediction capability of LLMs and serve as baselines for future studies.

GPT-4.0. A state-of-the-art commercial model from OpenAI (OpenAI, 2023). We choose the versions of `GPT-4-0314`.

GPT-3.5. To ensure reproducibility, we choose the `GPT-3.5-turbo-0301` a Snapshot of GPT-3.5-turbo (Ouyang et al., 2022) from March 1st, 2023.

GLM-130B. GLM-130B (Zeng et al., 2023) is an open bilingual dialog language model with 130 billion parameters and supports English and Chinese.

E.1 Experiment Settings

Following previous work (Shui et al., 2023), in the zero-shot setting, LLMs work following instructions without external law knowledge. In the few-shot setting, LLMs reason with prompts containing randomly selected (irrelevant) cases or similar cases retrieved by an information retrieval (IR) system. Fig. 10 shows the prompt template that is translated from Chinese. Since some fact descriptions are very long, we truncate them to 500 tokens. The Demo cases contain irrelevant cases or similar

Datasets	Language	Source	Domain	Legal System	# Pair	# Charge	# Law Article	# Term of Penalty	# Defendants/ Case	Conviction Elements
ECHR	English	ECHR	Human Right	Civil-Law	584	-	3	-	-	✗
ECHR-Case	English	ECHR	Human Right	Civil-Law	11,478	-	66	-	-	✗
ECHR-Crystal-Ball	English	ECHR	Human Right	Civil-Law	11,532	-	14	-	-	✗
QAjudge-CJO	Chinese	CJO	Criminal	Civil-Law	1,007,744	98	99	11	1.0	✗
Ajudge-PKU	Chinese	PKU	Criminal	Civil-Law	17,5744	68	64	11	1.0	✗
QAjudge-CAIL	Chinese	CAIL	Criminal	Civil-Law	113,536	105	122	11	1.0	✗
CAIL2018	Chinese	CJO	Criminal	Civil-Law	2,676,075	202	183	202	1.0	✗
CAIL-Long	Chinese	CJO	Criminal&Civil	Civil-Law	2,228,658	201	244	240	-	✗
Criminal-S	Chinese	CJO	Criminal	Civil-Law	61,589	149	-	-	1.0	✗
Criminal-M	Chinese	CJO	Criminal	Civil-Law	153,521	149	-	-	1.0	✗
Criminal-L	Chinese	CJO	Criminal	Civil-Law	306,900	149	-	-	1.0	✗
FLA	Chinese	CJO	Criminal	Civil-Law	60,000	50	-	-	1.0	✗
RACP	Chinese	CJO	Criminal	Civil-Law	100,000	50	-	-	1.0	✗
MultiLJP	Chinese	CJO	Criminal	Civil-Law	23,717	23	22	11	3.4	✗
ACI	English	SCI	Criminal	Common-Law	4,338	20	-	-	-	✗
MLMN	Chinese	CJO	Criminal	Civil-Law	1,189	2	86	-	-	✗
MUD	**Chinese**	**CJO**	**Criminal**	**Civil-Law**	**7,128**	**22**	-	-	**2.5**	✓

Table 7: A survey of datasets for charge prediction and related tasks. "#" denotes "the number of". "# Pair" denotes the number of Charge-Defendant pairs. CJO denotes China Judgment Online, PKU denotes Peking University Law Online, ECHR denotes the European Court of Human Rights, SCI denotes the Supreme Court of India.

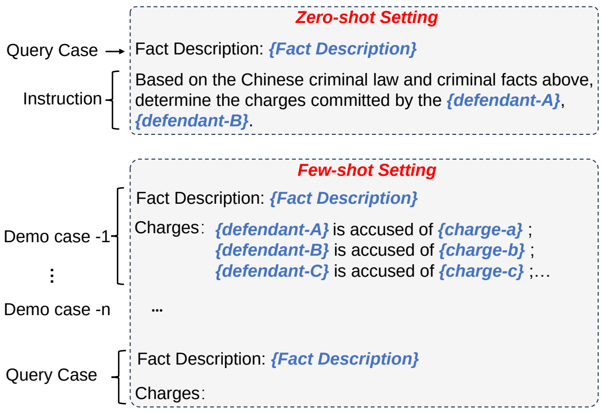

Figure 10: The prompt template translated from Chinese for zero- and few-shot charge prediction.

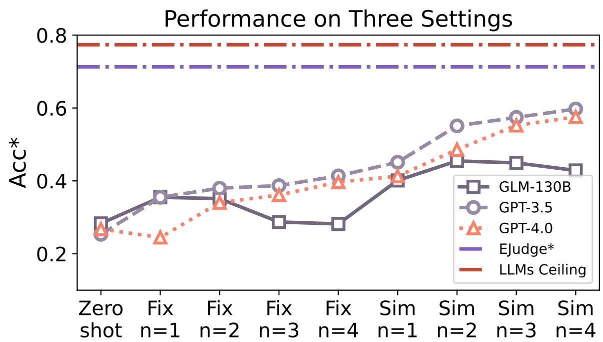

Figure 11: Results of LLMs on MUD, where "Zero shot", "Fix n", and "Sim n" represent prompting LLMs with instruction, fixed (irrelevant) n cases, and retrieved similar n cases, respectively.

cases. Specifically, for irrelevant cases, we randomly select several cases from the training dataset. For similar cases, we use the BM25 [3] algorithm to measure the similarity between the query case and cases in the training dataset, and top-n cases are kept.

E.2 Anasysis and Discusion

Fig. 11 shows the automatic evaluation result of LLMs.

For each LLM, few-shot baselines outperform zero-shot baselines, which conforms to our expectations. For few-shot baselines, LLMs prompting similar cases outperform LLMs prompting fixed cases, this is probably because the former import limit law knowledge compared to the the latter.

For GLM-130B, more similar cases or fixed cases in demonstrations are not always better. This is usually attributed to the noise introduced by irrelevant or false similar cases. GPT-4.0 and GPT-3.5 are more robust than GLM-130B.

It is slightly strange that LLMs perform worse than other baselines, such as Lawformer. This may be because LLMs can easily predict charges for the whole fact (LLM Ceiling in Fig. 11), but fail to align the charge for each defendant.

[3] https://pypi.org/project/rank-bm25/

指控名称(**Charges**)	定义(**Definitions**)
非法制造枪支罪(**Offences of Illegal Manufacture of Firearms**)	行为人违反国家有关枪支管理的法规，非法制造枪支、危害公共安全的行为。(**The perpetrator violated state regulations on firearms management by illegally manufacturing firearms and endangering public safety.**)
非法买卖枪支罪(**Offences of Illegal Trade in Firearms**)	行为人违反国家有关枪支管理的法规，非法买卖枪支、危害公共安全的行为。(**The perpetrator violated state regulations on firearms management by illegally trading in firearms and endangering public safety.**)
非法持有枪支罪(**Offences of Illegal Possession of Firearms**)	违反枪支管理规定,未经许可,非法持有枪支的行为。(**Illegal possession of firearms without authorisation in violation of firearms regulations.**)
销售假冒注册商标的商品罪(**Offence of Selling Counterfeit Registered Goods**)	销售明知是假冒注册商标的商品,销售金额较大的行为。(**Selling goods that are known to be counterfeit registered trademarks and selling a large amount of them.**)
合同诈骗罪(**Contract Fraud**)	以非法占有为目的，在签订、履行合同过程中，实施虚构事实或者隐瞒真相等欺骗手段，骗取对方当事人的财物，数额较大的行为。(**With the purpose of illegal possession, in the process of signing or fulfilling a contract, committing deceptive means such as fictitious facts or concealing the truth, to cheat the other party of property in a large amount.**)
非法经营罪(**Offence of Illegal Business Operation**)	违反国家规定,非法从事经营活动,扰乱市场秩序,情节严重的行为。(**Illegally engaging in business activities in violation of State regulations, disrupting the market order, under serious circumstances.**)
假冒注册商标罪(**Offence of Counterfeiting a Registered Trademark**)	违反国家商标管理法规，未经注册商标所有人许可，在同一种商品、服务上使用与其注册商标相同的商标，情节严重的行为。(**Violation of national trademark management regulations, without the permission of the owner of the registered trademark, in the same kind of goods and services, the use of the same trademark with its registered trademark, the circumstances are serious.**)
故意杀人罪(**Intentional Homicide**)	故意非法剥夺他人生命的行为。(**Intentional and unlawful deprivation of life.**)
故意伤害罪(**Intentional Injury**)	故意非法损害他人身体健康的行为。(**Acts of intentional unlawful damage to the physical integrity of another person.**)
非法拘禁罪(**Crime of Illegal Detention**)	故意非法拘禁他人或者以其他方法非法剥夺他人人身自由的行为。(**Deliberate unlawful detention of a person or other unlawful deprivation of a person's personal liberty.**)
抢劫罪(**Robbery**)	以非法占有为目的,使用暴力、胁迫或者其他方法,迫使被害人当场交出财物或者强行将公私财物当场抢走的行为。(**Using violence, coercion or other methods to force the victim to hand over property on the spot, or forcibly snatching public or private property on the spot, for the purpose of unlawful appropriation..**)
诈骗罪(**Fraud**)	以非法占有为目的,用虚构事实或者隐瞒真相的方法,骗取数额较大的公私财物的行为。(**Fraudulently obtaining a larger amount of public or private property by means of fictitious facts or concealment of the truth for the purpose of unlawful appropriation.**)
敲诈勒索罪(**Extortion and Blackmail**)	以非法占有为目的,对财物的所有人、管理人实施恐吓、威胁或者要挟的方法,强行索取数额较大的公私财物的行为。(**Intimidating, threatening or blackmailing the owner or manager of property for the purpose of unlawful appropriation, and forcibly soliciting a larger amount of public or private property.**)
招摇撞骗罪(**Crime of Cheating and Bluffing**)	为谋取非法利益，假冒国家机关工作人员的身份或职称，进行诈骗，损害国家机关的威信及其正常活动的行为。(**Fraudulent impersonation of the identity or title of a staff member of a State organ for the purpose of obtaining unlawful benefits, to the detriment of the prestige of the State organ and its normal activities.**)
聚众斗殴罪(**Crime of Affray**)	聚集多人攻击对方身体或者相互攻击对方身体，扰乱公共秩序的行为。(**Gathering of a number of persons to attack each other physically or to attack each other physically in order to disturb public order.**)
寻衅滋事罪(**Crime of Picking Quarrels and Provoking Troubles**)	肆意挑衅，随意殴打、骚扰他人或任意损毁、占用公私财物等行为，或者在公共场所起哄闹事，造成了严重破坏社会秩序的损害结果的行为。(**Acts of wanton provocation, randomly beating or harassing others or arbitrarily destroying or occupying public or private property, or acts of disturbances in public places that result in damages that seriously disrupt the social order.**)
掩饰、隐瞒犯罪所得罪(**Concealment of Proceeds of Crime**)	明知是犯罪所得，而予以窝藏、转移、收购、代为销售或者以其他方法掩饰、隐瞒的行为。(**Concealing, transferring, acquiring, selling or otherwise disguising or concealing the proceeds of crime, knowing that they are proceeds of crime.**)
窝藏、包庇罪(**Harboring and Covering**)	明知是犯罪的人而为其提供隐藏处所、财物，帮助其逃匿或者以作假证明的方式掩盖其罪行的行为。(**Providing a place of concealment or property to a person who has committed a crime, knowing that he or she has done so, assisting him or her to escape or concealing his or her crime by means of false testimony.**)
组织卖淫罪(**Crime of Organisation of Prostitution**)	以招募、雇佣、引诱、容留等手段，纠集、控制多人从事卖淫的行为。(**Recruiting, hiring, inducing, accommodating, etc., to gather and control a number of persons for the purpose of engaging in prostitution.**)
协助组织卖淫罪(**Crime of Facilitating the Organisation of Prostitution**)	为他人实施组织卖淫的犯罪活动提供方便、创造条件、排除障碍的行为。(**Facilitating, creating conditions and removing obstacles for others to commit the offence of organising prostitution.**)
容留卖淫罪(**Crime of harboring prostitution**)	为他人卖淫提供场所的行为。(**Provision of premises for the prostitution of others.**)
介绍卖淫罪(**Crime of Procuring prostitution**)	为卖淫的人与嫖客牵线搭桥的行为。(**Acts of matchmaking between persons engaged in prostitution and their clients.**)

Table 8: Rule of charges according to Criminal Law of the People's Republic of China, the translated version is indicated in bold font.

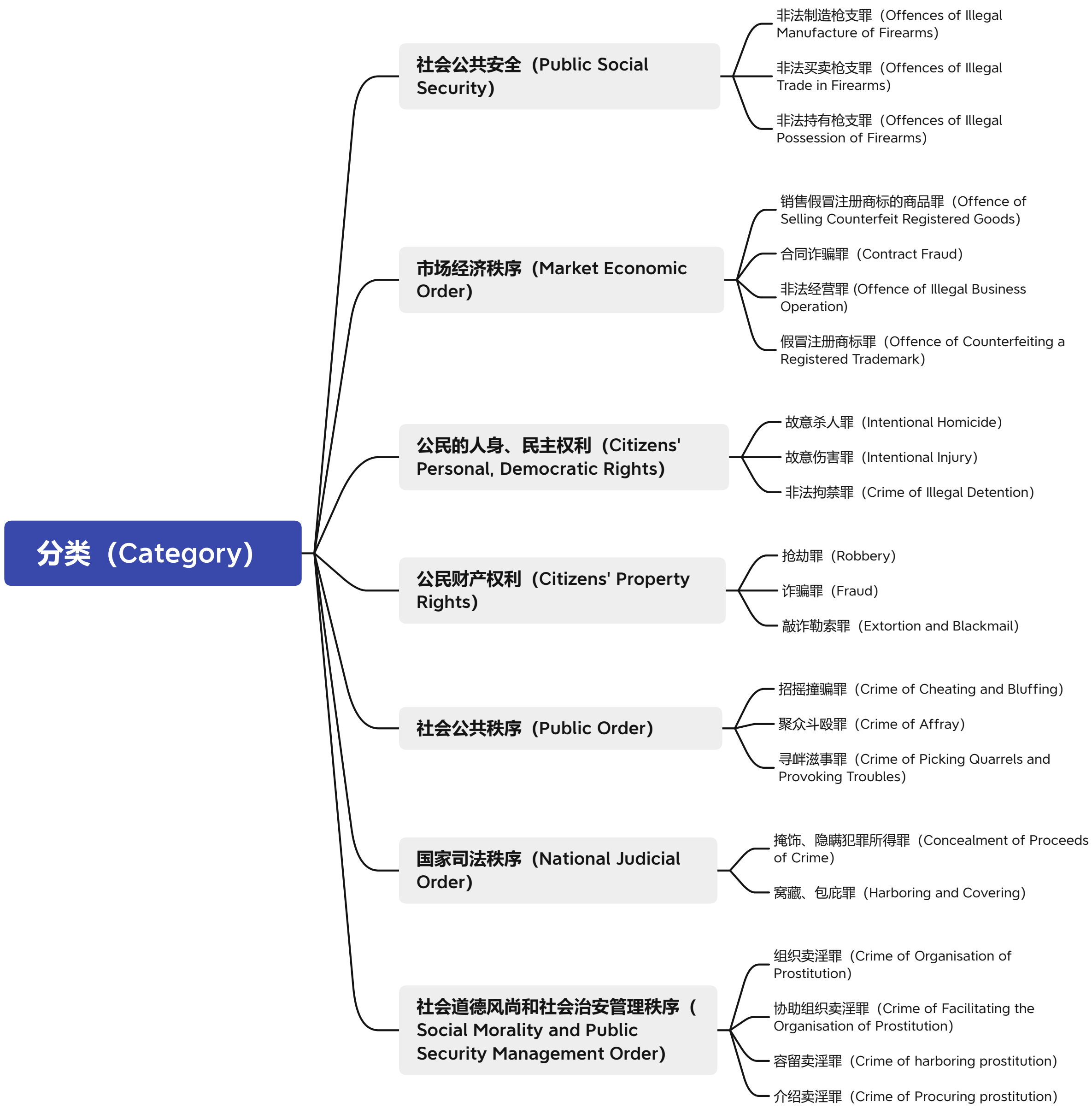

Figure 12: According to Criminal Law of the People's Republic of China, charges are arranged into different categories.