Expert Systems With Applications 249 (2024) 123626

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Knowledge-enhanced model with dual-graph interaction for confusing legal charge prediction[☆]

Sheng Bi [a,b], Zafar Ali [c], Tianxing Wu [c], Guilin Qi [c,*]

[a] *School of Law, Southeast University, Nanjing, China*
[b] *Judicial Big Data Research Centre, School of Law, Southeast University, Nanjing, China*
[c] *School of Computer Science and Engineering, Southeast University, Nanjing, China*

ARTICLE INFO

*Keywords:*
Confusing charge
Legal schematic knowledge
Dual graph
Knowledge matching network

ABSTRACT

The rapid development of natural language processing (NLP) technologies has enabled the emergence of legal intelligence assistance systems, with legal charge prediction (LCP) being a critical technology. The automatic LCP aims to determine the final charges based on fact descriptions of criminal cases. LCP assists human judges in managing workloads and improving efficiency, provides accessible legal guidance for individuals, and supports enterprises in litigation financing and compliance monitoring. However, distinguishing between confusing charges in real-world judicial practice remains a significant challenge. Most exist works cannot effectively capture complex relationships and discern subtle differences in fact descriptions while ignoring the legal schematic knowledge. In order to improve confusing LCP performance, we propose a novel knowledge-aware model for legal charge prediction that leverages Graph Neural Networks (GNNs) to capture complex relationships within criminal case descriptions. Specifically, the model constructs structural and semantic graphs from fact descriptions and integrates information from both through a dual-graph interaction process. A legal knowledge transformer generates key knowledge representations at schema and charge levels, while a knowledge matching network incorporates hierarchical charge knowledge into facts. Besides, we also propose two real-world datasets namely Criminal-All and Criminal-Confusing, containing 203 different charges and 86 confusing charges, respectively. To the best of our knowledge, these datasets are the first well-organized datasets for confusing LCP task. Experimental results demonstrate that the proposed model outperforms baselines and significantly improves the distinction of confusing charges, providing valuable support for intelligent legal judgment systems.

## 1. Introduction

In recent years, many fields have benefited from natural language processing (NLP) technologies thanks to the rapid development of data-driven and deep learning approaches (Aldunate, Maldonado, Vairetti, & Armelini, 2022; Araci, 2019; Chary, Parikh, Manini, Boyer, & Radeos, 2019). One of these applications is legal intelligence assistance systems (Surden, 2019; Zhong, Xiao, et al., 2020). The task of automatic charge prediction is to enable machines to identify suitable charges for a given criminal case description, such as **fraud** and **robbery** (Aletras, Tsarapatsanis, Preoţiuc-Pietro, & Lampos, 2016; Chalkidis, Androutsopoulos, & Aletras, 2019), which is a critical technology for intelligent legal judgment systems.

Legal Charge Prediction (LCP) can help human judges manage their workload and improve work efficiency. Individuals without legal background can consult machine judges for legal advice by describing their concerns and receive low-cost, high-quality legal assistance. From a management and business strategy perspective, companies can fund litigation costs for plaintiffs in exchange for a share of the proceeds if the case is successful (Armour & Dicker, 2019). LCP advances litigation finance by enabling investors to make more sophisticated, data-driven judgments about which cases are worth supporting. Law firms can use LCP to proactively plan litigation strategies, quickly track negotiations, and minimize the number of cases that require actual trials (Ahmed et al., 2018). In addition, companies can use LCP models to monitor their compliance with relevant laws and regulations. By analyzing internal documents, communications, and processes, the system can help identify potential violations and areas where the company needs to improve its compliance work (Maia, 2021).

Automatic LCP has been studied for many years and most existing work treats charge prediction as a text classification task (Liu, Chang,

[☆] We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

[*] Corresponding author.
*E-mail addresses:* bisheng@seu.edu.cn (S. Bi), zafarali@seu.edu.cn (Z. Ali), wutianxing@seu.edu.cn (T. Wu), gqi@seu.edu.cn (G. Qi).

https://doi.org/10.1016/j.eswa.2024.123626
Received 18 July 2021; Received in revised form 6 February 2024; Accepted 6 March 2024
Available online 8 March 2024

& Ho, 2004; Liu & Hsieh, 2006; Luo, Feng, Xu, Zhang, & Zhao, 2017; Wei & Lin, 2019). Early efforts extracted shallow text features, such as characters, words, and phrases, and used statistical learning algorithms such as support vector machines and random forests for prediction (Liu et al., 2004; Liu & Hsieh, 2006; Rosili et al., 2021; Sulea, Zampieri, et al., 2017). However, designing and extracting useful features is labor-intensive and time-consuming. Inspired by the success of deep neural networks in natural language processing tasks, numerous deep learning-based methods have been explored for charge prediction. In this context, Luo et al. (2017) proposed an attention-based neural model that supports charge prediction by selecting the most relevant legal provisions for the charge. As for multi-label charge prediction, Wei and Lin (2019) introduced an external knowledge-enhanced multi-label charge prediction model that combines legal articles with a deep learning network that can automatically adjust thresholds to obtain the number of legal charge labels. Ahmad, Asghar, Alotaibi, and Al-Otaibi (2022) suggested employing a hybrid deep learning-based decision support system, specifically a convolutional neural network (CNN) with bidirectional long short-term memory (BiLSTM), to predict court judgments. Some researchers have also combined several subtasks in legal judgment prediction into multi-task learning, believing that there us an inherent connection between these subtasks (Wu, Pan, Chen, Long, Zhang, & Yu, 2021). Nevertheless, the LCP presents significantly greater challenges and ambiguity than ordinary text classification. In real-world judicial practice, many confusing accusations exist, such as <*theft*, *robbery*, *defraud*>. The fact descriptions of these confusing charges have only subtle differences, which are difficult to capture. For instance, a *robbery* case may also encompass facts related to *theft*. The key distinction between these two accusations lies in whether the defendant had a subjective intention to harm the victim.

To address this issue, some researchers have begun to consider incorporating external information. Hu, Li, Tu, Liu, and Sun (2018) introduced several discriminative attributes for accusations to alleviate the few-shot charge predictions. They constructed ten discriminative legal attributes as internal mappings between fact descriptions and charges, providing signals to distinguish between confusing accusations. However, their method's drawback is the heavy reliance on experts, as summarizing and annotating attributes require substantial manual labor. Xu et al. (2020) proposed an end-to-end model to automatically learn the subtle differences between confusing legal articles and designed an attention mechanism to utilize the learned differences in extracting discriminative features from fact descriptions. Despite this, legal provisions do not contain enough information to differentiate between various accusations, especially those that are confusing. In practice, we hope for a method that can think like an expert in the legal domain, capable of learning and applying fundamental domain-specific knowledge to distinguish between different accusations.

From the description of criminal facts, it can be observed that discerning the complex relationships between different behaviors in the text is crucial to distinguishing confusing charges. These complex relationships involve both structural and semantic aspects and are hidden in long-distance texts. Graph Neural Networks (GNNs) have been proven to effectively model complex dependencies in long-distance sequences (Yao, Mao, & Luo, 2019). GNNs are deep learning models designed for graph-structured data, involving three key steps: initializing node representations, iteratively updating these representations through message passing, and combining the final representations using a readout function to accomplish tasks such as classification. GNNs have been successfully applied to text classification (Liang, Su, Gui, Cambria, & Xu, 2022; Liu, You, Zhang, Wu, & Lv, 2020; Ragesh, Sellamanickam, Iyer, Bairi, & Lingam, 2021). Empirical evidence suggests that GNNs possess advantages in capturing complex relationships, robustness to noise, and handling long-distance dependencies (Ragesh et al., 2021; Zhang, Tong, Xu, & Maciejewski, 2019).

Standing on the shoulders of our predecessors, we propose a novel knowledge-aware model to predict charges. In this model, we introduce legal schematic knowledge regarding criminal charges and utilize hierarchical knowledge representation as discriminative features to distinguish confusing charges. These features can provide explicit information on how to differentiate confusing charges. Specifically, our model takes textual fact descriptions as input and learns fact representations through GNN. To model the rich structural information and capture complicated semantic relationships, we construct a structural graph and a semantic graph from the fact descriptions, respectively. We then integrate the information from both graphs through a dual-graph interaction process to represent the facts. As a result, the fact representations encompass comprehensive global structural information and precise semantic relationships. Simultaneously, we employ a legal knowledge transformer to generate key knowledge representations oriented towards LCP at both the schema and charge levels. Furthermore, we apply a knowledge matching network to effectively incorporate hierarchical charge knowledge into facts, learning knowledge-aware fact representations. Finally, we use the knowledge-aware fact representations for charge prediction. To validate the effectiveness of our proposed model, we conduct a series of experiments on several real-world datasets. Comprehensive experimental results indicate that the proposed model outperforms other baselines and achieves significant improvements in terms of distinguishing confusing charges.

This paper highlights four main contributions to the automatic legal charge prediction:

1. We propose a novel knowledge-aware model that incorporates legal schematic knowledge related to criminal charges and uses hierarchical knowledge representation as discriminative features. This approach helps to more accurately discriminate between confusing charges and addresses the limitations of existing work.
2. We employ a dual-graph interaction process to integrate information from the structural and semantic graphs constructed from textual factual descriptions. This enables more comprehensive learning of subtle differences between confusing charges based on factual descriptions.
3. We implement a legal knowledge transformer to obtain schema- and charge-level-oriented representations from different hierarchies.
4. We introduce a deep matching network for facts and legal knowledge that effectively incorporates specific domain knowledge into the representation of criminal facts, resulting in more effective charge prediction.
5. Through a series of experiments on several real-world datasets, we demonstrate the effectiveness of our proposed model. The results indicate that the knowledge-aware model outperforms comparative methods and achieves a significant improvement in distinguishing between confusing charges.

## 2. Related work

This section is dedicated to discussing two principal topics related to our study: Legal Judgment Prediction (LJP) and Graph Neural Networks (GNNs). Our aim is to provide a thorough overview of their evolution.

### 2.1. Legal judgment prediction

Due to the fact that a large number of high-quality legal judgment documents have been made public, many researchers widely concern tasks for legal intelligence with utilizing NLP technology to help lawyers and other practitioners (Bi, Huang, Cheng, Wang, & Qi, 2019; Kien et al., 2020; Leitner, Rehm, & Moreno-Schneider, 2019; Shaikh, Sahu, & Anand, 2020; Zhong, Xiao, et al., 2020). The legal charge prediction task is the core of the legal intelligence system and has drawn increasing attention in recent years. In early studies on charge prediction, most researchers inclined to formalize it as a text classification problem (Aletras et al., 2016; Lin et al., 2012; Liu & Chen,

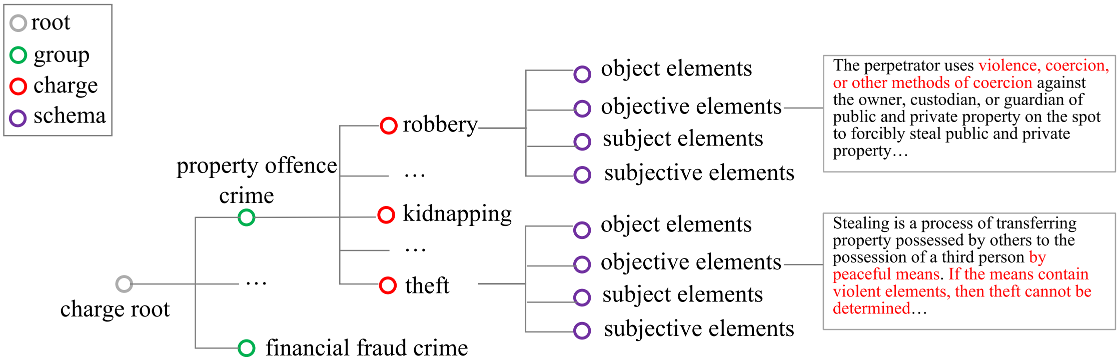

root
group
charge
schema
charge root
property offence crime
financial fraud crime
…
robbery
kidnapping
theft
object elements
objective elements
subject elements
subjective elements
The perpetrator uses violence, coercion, or other methods of coercion against the owner, custodian, or guardian of public and private property on the spot to forcibly steal public and private property…
Stealing is a process of transferring property possessed by others to the possession of a third person by peaceful means. If the means contain violent elements, then theft cannot be determined…

**Fig. 1.** An illustration of the legal schematic knowledge structure. The criminal law contains a variety of different groups, each group containing multiple charges. Each charge includes four elements, and different charges can be distinguished by constituting these elements.

2018; Liu & Liao, 2005; Marques, Bianco, Roodnejad, Baduel, & Berrou, 2019; Sulea, Zampieri, et al., 2017; Sulea, Zampieri, Vela & Van Genabith, 2017), which takes the fact description as an input and outputs a charge label by using machine learning. For instance, Bruninghaus and Ashley (2003) investigated an algorithm, Issue-Based Prediction (IBP), which combines reasoning with an abstract domain model and case-based reasoning techniques to predict the outcome of case-based legal arguments. Based on IBP, Ashley and Brüninghaus (2009) incorporated case-based reasoning and extracted information from fact description to predict and explain the outcomes of case scenarios. These classified cases are obtained by a set of classification concepts that capture stereotypical fact patterns that affect the strength of a legal claim. Liu et al. (2004), Liu and Hsieh (2006) used K-Nearest Neighbor (KNN) and extracted shallow textual features (e.g. characters, words, and phrases) artificially to predict charges. By taking data from the European Court of Human Rights as an example, Medvedeva, Vols, and Wieling (2019) addressed the potential in treating case law as quantitative data to predict judicial decisions and assessed how well Support Vector Machine (SVM) Linear Classifier is able to determine court judgments. However, these conventional methods can only leverage shallow textual features or manually designed factors, both need massive human efforts and hard to scale.

Motivated by the big success of deep learning in NLP tasks, researchers started to introduce neural network into charge prediction task (Chen, Cai, Dai, Dai, Ding, & Yadong, 2019; Li, Zhao, Li & Zhu, 2018; Shen et al., 2018; Undavia, Meyers, & Ortega, 2018; Wei, Qin, Ye, & Zhao, 2018; Yang, Wang, Zhang, Shou, Xu, & Wenwen, 2019). Given the fact description, Luo et al. (2017) proposed a hierarchical attention network for charge prediction by selecting the most relevant law articles to predict charges. Wei and Lin (2019) proposed a knowledge-aware end-to-end multi-label charge prediction method with an automatic label number learning network for multi-label charge prediction. Wei and Lin (2019) proposed a multi-label charge prediction method augmented by external knowledge, which is divided into two phases. One is a phase of charge label prediction using external knowledge from legal texts, and the other is a number learning phase. The method can automatically adjust the threshold to obtain the number of labels for legal cases by augmenting the external knowledge. It combines the output probabilities and their corresponding label numbers to obtain the final prediction results. However, these works fail to answer why the prediction results are correct, and the prediction results are hard to interpret. Therefore, Li, Zhang, Yu and Meng (2019) provided a cognitive computing framework for predicting judicial decisions, whose predicting results are interpretable in a way that induction rules are supplied. Zhong, Wang, et al. (2020) proposed a reinforcement learning based model to detect key elements in the fact description by iteratively asking questions, and then utilize the detected elements to predict judgment results.

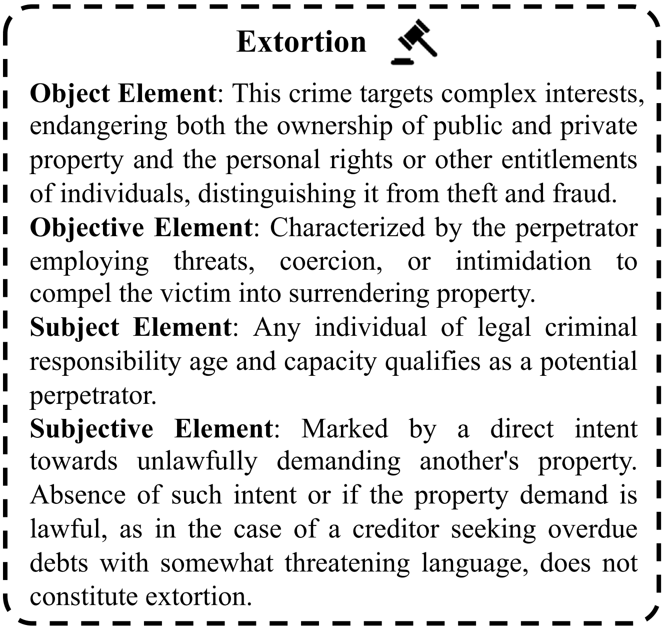
Nevertheless, all of these methods have poor performances on confusing charge prediction, which will influence the precise adjustment of prison term. Only few research works focus on confusing charges. Li, Liu, Ye, Zhang, Fang, and Binxing (2019) proposed an element-driven attentive neural network model, which introduces the legal constitutive elements as the discriminative features to distinguish confusing charges. To improve prediction accuracy, Yang, Jia et al. (2019) integrated word collocations features of fact descriptions into the network via an attention mechanism. These works do not consider the charge information, which is proved to be useful in predicting charges (Chen, Wang, Fang, Deng, Zhang, & Feng, 2019). Fortunately, Hu et al. (2018) constructed several discriminative attributes of charges as the internal mapping between fact descriptions and charges, which offer effective signals for distinguishing confusing charges. However, only a few confusing charges can be distinguished. It still remains a challenge to distinguish all confusing charges. Recently, Xu et al. (2020) focused on modeling subtle differences between confusing law articles and proposed a novel attention mechanism that exploits the learned differences to attentively extract discriminative features from fact descriptions for confusing charge prediction. Nevertheless, the information of law articles is not enough to distinguish different charges, especially the confusing charges. In practice, we would like to have a method that thinks like a legal domain expert and learns the elementary knowledge of how to distinguish different charges.

Therefore, the proposed knowledge-aware model aims to mitigate such problems by incorporating legal schematic knowledge about charges.

### 2.2. Graph neural networks

Graph Neural Networks (GNNs) are a class of deep learning models designed to handle graph-structured data. They are particularly effective in capturing complex relationships and dependencies within such data, which makes them well-suited for a variety of tasks, including text classification (Malekzadeh et al., 2021; Ragesh et al., 2021; Yao et al., 2019; Zhang et al., 2019). The basic idea of GNNs can be summarized in three steps: (1)Node representation: Each node in the graph represents an entity or a feature, and its initial representation can be obtained through various methods, such as embeddings or one-hot encoding. (2) Message passing: GNNs iteratively update node representations by aggregating information from their neighbors. This is done through a message-passing mechanism, where nodes exchange and combine information with their neighbors using a specific aggregation function. (3) Readout: After multiple iterations of message passing, the final node representations are combined through a readout function to form a graph-level representation, which can be used for downstream tasks like classification or regression.

Several variants of graph neural networks exist, such as GCN (Gao, Wang, & Ji, 2018; Zhang et al., 2019), GAT (Busbridge, Sherburn,



**Extortion**

**Object Element**: This crime targets complex interests, endangering both the ownership of public and private property and the personal rights or other entitlements of individuals, distinguishing it from theft and fraud.

**Objective Element**: Characterized by the perpetrator employing threats, coercion, or intimidation to compel the victim into surrendering property.

**Subject Element**: Any individual of legal criminal responsibility age and capacity qualifies as a potential perpetrator.

**Subjective Element**: Marked by a direct intent towards unlawfully demanding another's property. Absence of such intent or if the property demand is lawful, as in the case of a creditor seeking overdue debts with somewhat threatening language, does not constitute extortion.

**Fig. 2.** The constitutive components of **Extortion**.

Cavallo, & Hammerla, 2019; Veličković, Cucurull, Casanova, Romero, Lio, & Bengio, 2017; Wang et al., 2019), GraphSAGE (Hamilton, Ying, & Leskovec, 2017) and Gated GNN (Beck, Haffari, & Cohn, 2018; Li, Tarlow, Brockschmidt, & Zemel, 2015; Ruiz, Gama, & Ribeiro, 2020) which demonstrate the versatility and adaptability of GNNs in handling diverse graph-structured data and various learning tasks. GNNs, when dealing with text sequence style, typically model the text as a graph, where words are nodes and edges represent syntactic or semantic relationships. The model then learns to capture and propagate context information across the graph.

Empirical evidence suggests that Huang, Ma, Li, Zhang, and Wang (2019), Malekzadeh et al. (2021), Wu et al. (2019), GNNs are capable of capturing complex relationships between words in a text, which can lead to better understanding and representation of the text. Moreover, GNNs can capture long-range dependencies between words, even when the distance between them is large. This can help in understanding the context and capturing the overall meaning. On this basis, we design dual-graph interactions to integrate information from structural and semantic graphs constructed from textual fact descriptions. This allows us to learn more comprehensively the subtle differences between confusing charges.

## 3. Legal schematic knowledge

Fig. 1 outlines the overall framework of legal schematic knowledge, encompassing four hierarchical levels. Descending from the charge root node are various charge groups, serving as an aggregate categorization of all offenses. For instance, crimes of property offense and financial fraud are highlighted. Each group further branches into specific charges, where crimes within the same category may present similarities that could lead to confusion. It is noteworthy that the structure encompasses 25 groups encompassing 191 distinct charges. To facilitate distinguishing between crimes that could easily be confused, we employ the knowledge of their constitutive elements to define their distinct characteristics. Specifically, each charge is analyzed through four components: object elements $K_1$, objective elements $K_2$, subject elements $K_3$ and subjective elements $K_4$. These components serve as fundamental criteria for crime differentiation, providing a basis for clear delineation among similar offenses. Subsequently, we delve into the significance and functionality of these four components. Additionally, Fig. 2 visually represents the constitutive characteristics of "extortion", facilitating an intuitive comprehension of these four critical elements.

Under Chinese Criminal Law, the definitive features of each crime are segmented into four elemental paradigms, collectively delineating the legitimacy of a criminal act and its legal assessment. These elements are expounded as follows:

(1) Subjective Element: This pertains to the mental state of the perpetrator at the crime's commission, including intentions and negligence, further bifurcated into direct and indirect intentions, and negligence into careless and overconfident types. This element plays a pivotal role in establishing the occurrence of a crime. Notably, criminal law recognizes negligence, alongside intent, as a basis for constituting a crime.

(2) Objective Element: This element encapsulates the tangible manifestation of the criminal act, including its method, outcome, and the causality linking the act to its consequence. It serves as the external criterion for crime determination and the factual foundation for evaluating an act's criminality.

(3) Subject Element: This concerns the perpetrator of the criminal act, encompassing age and mental condition considerations. Only individuals of legal age and possessing criminal responsibility are deemed capable of criminal conduct, thus identifying liable parties.

(4) Object Element: This refers to the societal relationships or legal rights infringed upon by the criminal act, such as life, health, property, and public safety. It is a prerequisite for crime establishment, with each crime targeting specific legal interests.

These elements interrelate and restrict each other, forming the adjudication foundation of crimes within Chinese Criminal Law. A comprehensive analysis of these elements is imperative for accurately determining the nature of a criminal act and its classification.

## 4. Methodology

In this section, we delineate the intricacies of our model. The comprehensive architecture of our model is depicted in Fig. 3, encompassing three components, i.e., a fact encoder, a knowledge encoder, and a charge prediction module. The fact encoder ingests the textual fact description as input and constructs two graphs, concentrating on structural information and semantic relations correspondingly, executing dual-graph interaction to acquire information for prediction. Subsequently, the knowledge encoder employs the LK-Transformer to engender pivotal knowledge representations oriented towards the legal schematic knowledge at both schema and charge echelons. To augment the distinctiveness, the group label information is encoded and amalgamated with the corresponding knowledge representations. Ultimately, within the charge prediction module, we implement a knowledge matching network for efficaciously integrating charge information into the fact to ascertain knowledge-aware fact representation. Employing the knowledge-aware fact representation, we utilize a softmax layer to exhibit the anticipated distributions of charges.

In the ensuing subsections, we initially proffer the problem formulation and introduce our legal schematic knowledge. Following that, we expound upon the neural encoder of fact description and the knowledge-aware fact representation. Conclusively, we display the output layer and the loss function of our model.

### *4.1. Problem formulation*

For each criminal case, the fact description is considered a lexical concatenation $X = (x_1, x_2, \dots, x_n)$, where $n$ represents the length of $X$. We derive legal schematic knowledge $K$ concerning $m$ charges. Provided the fact description $X$ and legal schematic knowledge $K$, the legal charge prediction (LCP) endeavor aspires to foresee a charge label $y$. Table 1 exhibits the specific connotation of each parameter.

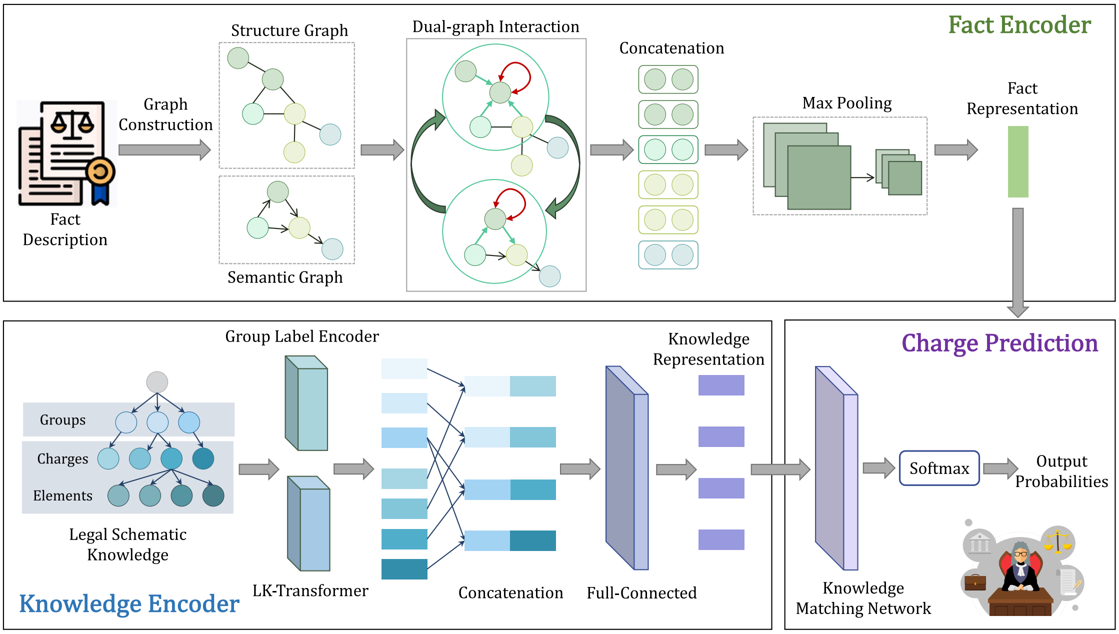


**Fig. 3.** An illustration of our proposed model for charge prediction. Our model consists of three parts, including a fact encoder, a knowledge encoder and a charge prediction module. The fact encoder constructs a semantic graph and a structure graph to obtain the fact representation by dual-graph interaction. Meanwhile, the knowledge encoder employs a LK-Transformer to generate crucial knowledge representations at both the schema and charge levels. Finally, take the fact and knowledge as inputs, the charge prediction module leverages a knowledge matching network to generate knowledge-aware fact representation and then predict charges.

**Table 1**
The meaning of parameters in problem formulation.

| Parameter | Meaning |
|---|---|
| $X$ | the fact description regraded as a word sequence |
| $x$ | a word in the fact description |
| $n$ | the number of words in the fact description |
| $K$ | the extracted legal schematic knowledge |
| $m$ | the number of charges |
| $y$ | a charge label to be predicted |

### 4.2. *Fact representation*

As shown in Fig. 3, the fact encoder encodes the discrete input sequence into hidden states, obtaining the fact representation through max pooling. Traditional neural encoders, such as RNNs and CNNs, effectively capture semantic and syntactic information in local consecutive word sequences, but ignore global word co-occurrence in a corpus, which conveys non-consecutive and long-distance semantics. To address this issue, we construct a structural graph $U$ using fact descriptions and apply a graph convolutional network (GCN) for encoding. The GCN, a straightforward yet efficient graph neural network, captures higher-order neighborhood information and rich global structural data. Furthermore, the extraction of semantic relationships between entities in fact descriptions assists in identifying charge-worthy content and determining charge types. Consequently, we also create a semantic graph $V$ to extract semantic relationships. We then integrate information from both $U$ and $V$ during a dual-graph interaction process to represent the fact. As a result, the fact representation encompasses not only abundant global structural information but also significant semantic relationships.

#### 4.2.1. *Structure graph construction*

First, the fact encoder transforms each word $x_i \in X$ into its corresponding word embedding $x_i \in R^d$ via an embedding layer, where $d$ represents the dimension of word embeddings. The resulting word embedding sequence, denoted as $W = (w_1, w_2, \ldots, w_n)$, serves as the initial feature matrix $H^{(0)}$. Subsequently, a structure graph is constructed for the given text, with all words $W = (w_1, w_2, \ldots, w_n)$ in the text considered as the graph nodes $U = (u_1, u_2, \ldots, u_n)$. To calculate edge weights, the point-wise mutual information (PMI) is utilized, effectively preserving global word co-occurrence information (Yao et al., 2019). Specifically, a fixed-size sliding window is applied to all documents in both the source and legal domains to gather word co-occurrence statistics. The PMI of a word pair $w_i$, $w_j$ is computed as:

$$p(w_i, w_j) = \frac{W(w_i, w_j)}{|W|}, \tag{1}$$

$$PMI(w_i, w_j) = log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{2}$$

where $p(w_i) = \frac{W(w_i)}{|W|}$. $W(w_i)$ denotes the count of sliding windows containing the word $w_i$. Additionally, $W(w_i, w_j)$ represents the number of sliding windows encompassing both words $w_i$ and $w_j$, while $|W|$ represents the total number of sliding windows. The PMI score can effectively capture the semantic relationship between words. A negative PMI value indicates minimal or no semantic correlation. Consequently, edges are added exclusively between word pairs exhibiting positive PMI scores:

$$a_{ij} = \begin{cases} PMI(w_i, w_j), & PMI(w_i, w_j) > 0, \\ 0, & PMI(w_i, w_j) \leq 0. \end{cases} \tag{3}$$

where $a_{ij}$ represents the relationship between words $w_i$ and $w_j$. By processing these relationships, we acquire the word relations $\mathcal{A}$ across the global corpus, with the adjacency matrix $A$ constituting a subset of $\mathcal{A}$ for each document.

#### 4.2.2. *Semantic graph construction*

To construct the semantic graph from fact descriptions, we employ Dependency Parsing (DP)-based techniques. DP aims to establish relationships between “head” characters and their modifying characters within a tree structure, conveying the sentence’s semantic information

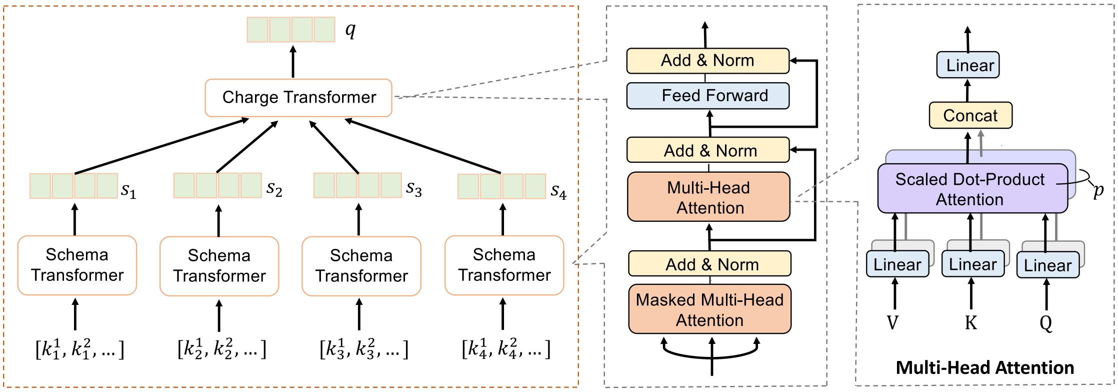


**Fig. 4.** An illustration of the LK-Transformer. The middle part is the explicit structure of transformer and the right part is the explanation of multi-head attention.

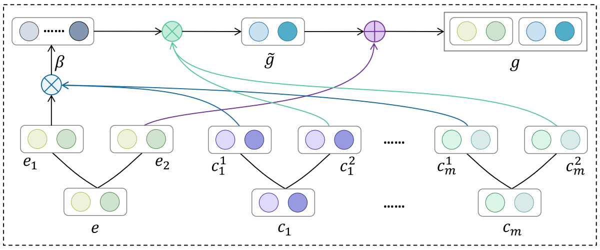


**Fig. 5.** An illustration of the knowledge matching network. $e$ denotes fact representation and $c_i$ indicates each knowledge word representation, where $i \in [1, s]$. ⊗ means the multiply operation and ⊕ means the concatenate operation.

through an analysis of its grammatical structure. As fact descriptions often contain numerous pronouns that may impede entity connections, we first use AllenNLP's coreference analysis system[1] to replace pronouns referring to the same entity with the original entity name. Subsequently, we utilize AllenNLP's implementation of the biaffine attention model (Dozat & Manning, 2017) to obtain the dependency parse tree $T_s$ for each sentence $s$. We then refine $T_s$ by removing unnecessary components (e.g., punctuation marks) and merging the continuous nodes that constitute a cohesive semantic unit. Lastly, we introduce inter-tree edges between similar nodes from distinct parse trees, thus creating a connected semantic graph $V$.

In the graph $V$, each node $v_i = \{\omega_{ij}\}_{j=m_{v_i}}^{n_{v_i}}$ represents a text span in the fact description, accompanied by a corresponding node type $t_v$. The starting and ending positions of the text span are denoted by $m_{v_i}$ and $n_{v_i}$, respectively. To convey intricate semantic relationships, we assign each edge a specific type $t_e$. As the majority of text spans comprise several words, we concatenate the embeddings of their constituent words to generate the initial node representations $h_v^0$.

### 4.2.3. Dual graph interaction

In our structural graph $U$, node representations encapsulate information corresponding to the global structure, whereas node representations in the semantic graph $V$ capture rich semantic information. To more effectively represent fact descriptions, it is intuitively beneficial to integrate information from both $U$ and $V$. Consequently, we propose a dual-graph interaction process wherein $U$ and $V$ iteratively update their node representations by sharing features with each other. This dual-graph process comprises three steps.

*Information propagation.* Upon constructing the two graphs and obtaining the initial representations, the first step entails updating each node's representation via message passing.

For the structural graph $U$, the propagation rule can be interpreted as Laplacian smoothing (Li, Han, & Wu, 2018). A node's new feature is calculated as the weighted average of its own and its neighbors' features, followed by a linear transformation. Additionally, each node can gather and integrate messages from adjacent nodes to update its representation. At time step $t$, the received message $h_{u_i}^t$ of node $u_i$ is defined as:

$$h_{u_i}^t = \sum_{u_j \in \mathcal{N}_{(u_i)}} W_{u_{ij}} u_j^{(t-1)} + b_{u_{ij}}, \tag{4}$$

where $\mathcal{N}_{(u_i)}$ represents the set of all neighbors of node $u_i$, while $W_{u_{ij}}$ and $b_{u_{ij}}$ are trainable parameters controlling message aggregation. For the semantic graph $V$, given that multiple semantic relations exist within the edges, we draw inspiration from the multi-relation Gated Graph Neural Network (GGNN) (Li et al., 2015; Ruiz et al., 2020) to obtain a separate transformation matrix for each edge type. Furthermore, since not all neighboring nodes are essential in message passing, we employ an attention mechanism to dynamically model the weights of neighboring nodes. More specifically, at each transition time step, each node $v_i$ receives information from neighboring nodes, which is defined as follows:

$$h_{\mathcal{N}_{\vdash(i)}}^{(t)} = \sum_{v_j \in \mathcal{N}_{\vdash(i)}} \alpha_{ij}^t W^{t_{e_{ij}}} h_j^{(t)}, \tag{5}$$

$$h_{\mathcal{N}_{\dashv(i)}}^{(t)} = \sum_{v_j \in \mathcal{N}_{\dashv(i)}} \alpha_{ij}^t W^{t_{e_{ji}}} h_j^{(t)}, \tag{6}$$

where $W^{t_{e_{ij}}}$ is the weight matrix corresponding to the edge type $t_{e_{ij}}$ from $v_i$ to $v_j$. $\mathcal{N}_{\vdash(i)}$ and $\mathcal{N}_{\dashv(i)}$ are the sets of incoming and outgoing edges of $v_i$, respectively. $\alpha_{ij}^t$ is the attention coefficients of $v_i$ over $v_j$, calculated as follows:

$$\alpha_{ij}^t = \frac{exp(Attn(W^A h_i^t, W^A h_j^t))}{\sum_{k \in \mathcal{N}_{\vdash(i)}} exp(Attn(W^A h_i^t, W^A h_k^t))}, \tag{7}$$

where $Attn$ represents a single-layer feed-forward neural network to perform attention on the nodes. $W^A$ are shared weight matrix for every node to perform linear transformation. Ultimately, the aggregated neighboring information $h_{v_i}^t$ is computed as:

$$h_{v_i}^t = [h_{\mathcal{N}_{\vdash(i)}}^{(t)}; h_{\mathcal{N}_{\dashv(i)}}^{(t)}] \tag{8}$$

*Gates calculation.* In the second step, we calculate two gates, including an update gate $z_{u_i}^t$ and a reset gate $r_{u_i}^t$. The update gate and the reset gate are used to decide how much information should be passed and forgotten at each time step in the propagation process, respectively. $z_{u_i}^t$ and $r_{u_i}^t$ are defined by:

$$z_{u_i}^t = \sigma(W_{z_u}[h_{u_i}^t; u_i^{t-1}]), \tag{9}$$

[1] https://allenai.org/allennlp

$$r_{u_i}^t = \sigma(W_{r_u}[h_{u_i}^t; u_i^{t-1}]), \tag{10}$$

where $W_{z_u}$ and $W_{r_u}$ are parameters. For semantic graph $V$, we similarly compute:

$$z_{v_i}^t = \sigma(W_{z_v}[h_{v_i}^t; v_i^{t-1}]), \tag{11}$$

$$r_{v_i}^t = \sigma(W_{r_v}[h_{v_i}^t; v_i^{t-1}]). \tag{12}$$

*Information interaction.* In order to facilitate feature sharing and information interaction, we employ update and reset gates derived from one graph to modify node representations within each respective graph. Given that graph $U$ considers individual words as nodes and graph $V$ treats text spans as nodes, we adopt the following approach: For graph $U$, we utilize the corresponding text span gates to refine the word representation. For graph $V$, we compute the average value of the gates corresponding to its constituent words in order to update the text span representation. Concretely, new node representations are defined by:

$$\begin{aligned} u_i^t =& r_{v_j}^t \odot u_i^{t-1} + (1 - r_{v_j}^t) \\ &\odot tanh(W_{u_h}[h_{u_i}^t; z_{v_j}^t \odot u_i^{t-1}]), \text{ if } u_i \in v_j, \\ v_j^t =& \sum_{u_i \in v_j} r_{u_i}^t \odot v_j^{t-1} + (1 - \sum_{u_i \in v_j} r_{u_i}^t) \\ &\odot tanh(W_{v_h}[h_{v_j}^t; \sum_{u_i \in v_j} z_{u_i}^t \odot v_j^{t-1}]). \end{aligned} \tag{13}$$

After executing the three steps for $T$ times iteratively, we obtain the representations $u_i^{(T)}$ and $v_j^{(T)}$ for $u_i \in U$ and $v_j \in V$. We concatenate each word representation $u_i^{(T)}$ with its corresponding text span representation $v_j^{(T)}$ to get the semantic-enriched word representations, which are then input to a max-pooling layer to obtain the final fact representation $\vec{e} = [e_1, \dots, e_s]$ as

$$u_i^{(T)} = [u_i^{(T)}; v_j^{(T)}], \text{if } u_i^{(T)} \in v_j^{(T)}, \tag{14}$$

$$e_i = \max(u_{1,i}^{(T)}, \dots, u_{n,i}^{(T)}), \forall i \in [1, s], \tag{15}$$

here, $s$ is the dimension of hidden states.

### 4.3. Knowledge-aware fact representation

In order to address the challenge of distinguishing between confusing charges, we leverage the legal schematic knowledge representation of charges as discriminative features. As previously mentioned, the availability of legal schematic knowledge enables us to identify subtle differences in specific aspects of criminal motive, action, or consequence among these charges and subsequently differentiate them.

#### 4.3.1. LK-transformer

Owing to the tree structure of our legal schematic knowledge, we employ the LK-Transformer to generate essential knowledge representations at both schema and charge levels. The architecture of our LK-Transformer is depicted in Fig. 4. For each charge, we input the schemas $K_1$, $K_2$, $K_3$, and $K_4$ into four schema transformers, respectively, and obtain corresponding semantic representations $s_1$, $s_2$, $s_3$, and $s_4$. Subsequently, these four schema representations are input into a charge-level transformer to compute the knowledge representation $q$.

More specifically, both the schema-level and charge-level transformers utilize the same model, as illustrated in the central portion of Fig. 4. The model comprises three components: a masked multi-head self-attention mechanism, a multi-head self-attention mechanism, and a position-wise fully connected feed-forward network. We implement a residual connection (He, Zhang, Ren, & Sun, 2016) around each of these three sub-layers, followed by layer normalization (Ba, Kiros, & Hinton, 2016). Thus, the output of each sub-layer is expressed as $LayerNorm(x + Sublayer(x))$, where $Sublayer(x)$ represents the function executed by the sub-layer itself. To support these residual connections, all sub-layers and embedding layers within the model yield outputs with a dimension of $d_{model} = 512$.

*Multi-head attention.* The attention mechanism maps a query and a set of key–value pairs to an output in a soft manner, with the query, keys, values, and output all represented as vectors. Accepting queries and keys of dimension $d_k$, and values of dimension $d_v$, it produces a weighted sum of the values. In the case of multi-head attention, multiple individual attention functions operate in parallel, generating $d_v$-dimensional output values. These are concatenated and subsequently projected to produce the final values, as illustrated in the right half of Fig. 4. Multi-head attention is advantageous because it enables joint attention to information from distinct representation subspaces at varying positions, whereas a single attention head would be constrained by averaging. Instead of employing a single attention function,Vaswani et al. (2017) discovered that utilizing multiple individual attention functions is beneficial for capturing different contexts.

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, \dots, head_h)W^O, \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V), \end{aligned} \tag{16}$$

where the projections are parameter matrices $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_v}$ and $W^O \in R^{pd_v \times d_{model}}$. In this work we employ $p = 8$ parallel attention layers, or heads. For each of these we use $d_k = d_v = d_{model}/p = 64$.

*Position-wise feed-forward networks.* Besides attention sub-layers, each of the layers has a fully connected feed-forward network, which consists of two linear transformations with a ReLU activation (Nair & Hinton, 2010).

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2. \tag{17}$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer.

#### 4.3.2. Knowledge–aware fact representation

Through the LK-Transformer, we obtain knowledge representations. To enhance their distinctiveness, we implement a BiLSTM (Zhang, Zheng, Hu, & Yang, 2015) to encode group label information. The representations are then concatenated with the corresponding knowledge representations to differentiate between charges, yielding the final knowledge representations $\boldsymbol{c} = (c_1, c_2, \dots, c_m)$.

We utilize a knowledge matching network to select relevant information from knowledge, generating a knowledge-aware fact representation. Let $\boldsymbol{e_j} = W_j^e \boldsymbol{e}$ and $c_i^j = W_{ij}^c c_i$. As depicted in Fig. 5, the knowledge matching network initially computes the attention $\boldsymbol{\beta}$ for fact representation $\boldsymbol{e_1}$ and knowledge representation $\boldsymbol{c^1}$. Subsequently, the matched knowledge embeddings $\tilde{\boldsymbol{g}}$ for the fact are acquired by multiplying the attention matrix $\boldsymbol{\beta}$ with the knowledge representation $\boldsymbol{c^2}$. Lastly, we concatenate the matched knowledge embeddings with the fact representation $\boldsymbol{e_2}$ to obtain the knowledge-aware fact representation $\boldsymbol{g}$.

Our knowledge matching function is defined as following:

$$\begin{aligned} \boldsymbol{g} &= \mathcal{F}(\boldsymbol{e_1}, \boldsymbol{c^1}, \boldsymbol{e_2}, \boldsymbol{c^2}) \\ &= Concat(\boldsymbol{e_2}; \tilde{\boldsymbol{g}}) \\ &= Concat(\boldsymbol{e_2}; \boldsymbol{c^2}\boldsymbol{\beta}^T), \end{aligned} \tag{18}$$

where the function Concat is a concatenation operation, and $\beta$ is an attention score, computed by:

$$\boldsymbol{\beta} \propto exp(ReLU(W\boldsymbol{e_1})^T ReLU(W\boldsymbol{c^1})), \tag{19}$$

where $W$ is a learnable weight matrix and ReLU is the rectified linear unit.

### 4.4. Prediction and optimization

Finally, we employ the knowledge-aware fact representation $\boldsymbol{g}$ to predict a case's final charge in the output layer. The predicted probability distribution $y$ across all charges is computed as $y = \text{softmax}(W^y \boldsymbol{g} +$

**Table 2**
The detailed information of our constructed datasets.

| | Judgments | Charges | Only confusing charges |
|---|---|---|---|
| Criminal-All[a] | 336 450 | 203 | ✗ |
| Criminal-Confusing[b] | 80 000 | 86 | ✓ |

[a] https://github.com/thunlp/CAIL2018

[b] https://github.com/ai4law/legaldata

$b^y$), where $W^y$ and $b^y$ denote the weight matrix and bias vector in the output layer, respectively. Our model's training objective is to minimize the cross-entropy between the predicted charge $\hat{y}$ and the ground-truth $y$. The charge prediction loss can be formalized as:

$$Loss = -\sum_{i=1}^{m} y_i \cdot log(\hat{y}_i). \tag{20}$$

## 5. Experiments

In this section, we validate the efficacy of our model for criminal charge prediction through a series of experiments conducted on two real-world datasets, comparing our proposed model with several state-of-the-art baselines.

### 5.1. Dataset construction

Table 2 provides detailed information on our constructed datasets. We have gathered 336,450 criminal case judgments published by CAIL2018 (Xiao et al., 2018). Original legal judgment comprises several components (Bi, Ali, Wang, Wu, & Qi, 2022), such as defendant information, fact description, and charges. We retain only the fact description as input and the charge as the label. The dataset, referred to as Criminal-All, encompasses 203 distinct charges. Furthermore, we select 80,000 cases involving 86 confusing charges to create another dataset, termed Criminal-Confusing, which is designed to test our model's ability to predict confusing charges. For both Criminal-All and Criminal-Confusing, we randomly allocate 80% of the cases for training, 10% for validation, and 10% for testing.

### 5.2. Baselines

To evaluate the performance of our framework, we compare our method with the following baselines:

* **HAN**: a Hierarchical Attention based RNN for document classification (Yang et al., 2016). We set the word embedding dimension to be 200, the GRU dimension to be 50, the batch size to be 64 and the learning rate to be 0.001.
* **DPCNN**: a low-complexity word-level deep convolutional neural network architecture for text categorization (Johnson & Zhang, 2017). We set the word embedding dimension to be 250, the batch size to be 100 and the learning rate to be 0.001.
* **Few-shot**: an attribute-based multi-task learning model for charge prediction by introducing discriminative legal attributes into consideration (Hu et al., 2018). We set the word embedding dimension to be 100, the batch size to be 64 and the learning rate to be 0.001.
* **EDA-NN**: an element-driven attentive neural network model which can jointly predict the legal constitutive elements and judgment results (Li, Liu, et al., 2019). We set the word embedding dimension to be 100, the GRU dimension to be 100, the batch size to be 64 and the learning rate to be 0.001.
* **MPBN**: a multi-perspective bi-Feedback network with the word collocation attention mechanism for confusing charge prediction (Yang, Jia et al., 2019). We set the word embedding dimension to be 200, the hidden state to be 256, the batch size to be 128 and the learning rate to be 0.001.

**Table 3**
Charge prediction results of two datasets.

| Datasets | Criminal-All | | | | Criminal-Confusing | | | |
|---|---|---|---|---|---|---|---|---|
| Metrics | Acc. | MR | Ma-F1 | Mi-F1 | Acc. | MR | Ma-F1 | Mi-F1 |
| HAN | 0.858 | 0.710 | 0.727 | 0.673 | 0.761 | 0.612 | 0.631 | 0.582 |
| DPCNN | 0.878 | 0.767 | 0.783 | 0.719 | 0.782 | 0.653 | 0.675 | 0.614 |
| Few-shot | 0.891 | 0.824 | 0.842 | 0.788 | 0.844 | 0.783 | 0.798 | 0.762 |
| EDA-NN | 0.901 | 0.843 | 0.866 | 0.803 | 0.863 | 0.801 | 0.814 | 0.787 |
| MPBN | 0.903 | 0.858 | 0.875 | 0.810 | 0.870 | 0.811 | 0.826 | 0.792 |
| LADAN | 0.905 | 0.862 | 0.879 | 0.815 | 0.875 | 0.817 | 0.829 | 0.796 |
| **Ours** | **0.924** | **0.901** | **0.914** | **0.857** | **0.910** | **0.878** | **0.889** | **0.824** |

* **LADAN**: a Law Article Distillation based Attention Network to distinguish confusing charges (Xu et al., 2020). We set the hidden state to be 256, the batch size to be 128 and the learning rate to be 0.001.

Note that all parameters of the baseline models are selected for their optimal performance. For the evaluation, we employ accuracy (Acc.), macro-recall (MR), macro-F1 (Ma-F1) and micro-F1 (Mi-F1) as metrics.

### 5.3. Experimental settings

Initially, we employ jieba[2] for Chinese word segmentation and set the maximum document length to 500. Subsequently, we utilize word2vec (Mikolov, Chen, Corrado, & Dean, 2013) to train word embeddings on all legal judgments, with an embedding size of 300. Regarding hyperparameter settings, we establish the dimension of all hidden states as 256. For training, we use Adam (Kingma & Ba, 2014) as the optimizer, set the learning rate to 0.001, and configure the batch size and dropout rate as 64 and 0.5, respectively. We continue training iterations until the difference between two consecutive iterations is sufficiently small.

### 5.4. Results and discussion

As illustrated in Table 3, our model outperforms all baseline models on Criminal-All, demonstrating the robustness and effectiveness of our proposed method for charge prediction. HAN and DPCNN, general text classification models with limited scalability, perform poorly on specific domains such as legal charge prediction tasks. Few-shot and EDA-NN construct only several discriminative attributes of charges, providing insufficient and ineffective signals for distinguishing charges. MPBN incorporates word collocation features of fact descriptions but neglects charge discriminative attributes. Among these baselines, LADAN achieves the best performance as it encodes law article information as discriminative attributes. However, law articles alone are insufficient for distinguishing different charges, particularly confusing charges, compared to legal schematic knowledge. Consequently, our model outperforms LADAN.

Generally, the performance of all the methods is better on dataset Criminal-All compared to dataset Criminal-Confusing. This could be attributed to the larger size and more diverse categories in Criminal-All compared to Criminal-Confusing. The larger dataset might provide more diverse and representative samples, enabling the models to learn better representations and make better predictions. The gap in performance between the proposed method and the other methods is more pronounced for the macro metrics (MR, Ma-F1) than for the micro metrics (Acc, Mi-F1). This suggests that the proposed method might be particularly effective at handling class imbalance or improving performance on underrepresented categories.

To further assess our model's effectiveness in handling confusing charges, we present its performance on the Criminal-Confusing dataset

[2] https://github.com/fxsjy/jieba

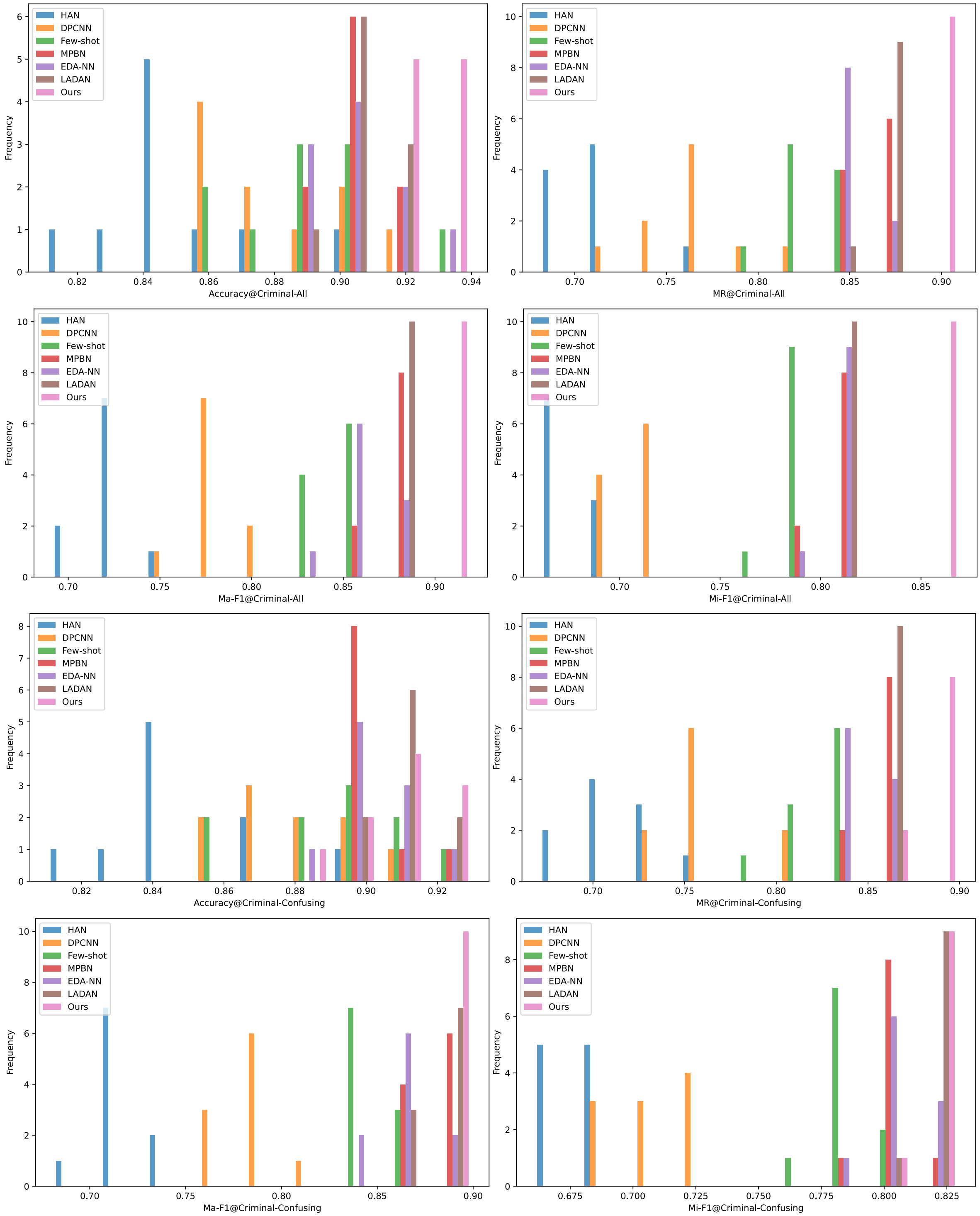


**Fig. 6.** The results of the ten-fold cross-validation on Criminal-All and Criminal-Confusing.

in Table 3, comparing it with LADAN, the state-of-the-art model for predicting confusing charges. Our model improves the performance by approximately 2.3%, 2.4%, and 2.7% relatively on accuracy (Acc.), macro recall (MR), and macro F1-score (Ma-F1), respectively, which highlights our model's capabilities.

### 5.5. Cross validation

In this section, we present a detailed plan for applying ten-fold cross-validation to evaluate the performance of our proposed model, which will enable us to draw robust and reliable conclusions from our experiments.

To perform ten-fold cross-validation, we first need to divide the dataset into ten equally-sized folds.[3] Each fold will serve as a test set for one iteration, while the remaining nine folds will be used as the training set. In our experiment, we will use a stratified approach to

[3] For simplicity, the validation set is fixed and only the training and test sets are rearranged.

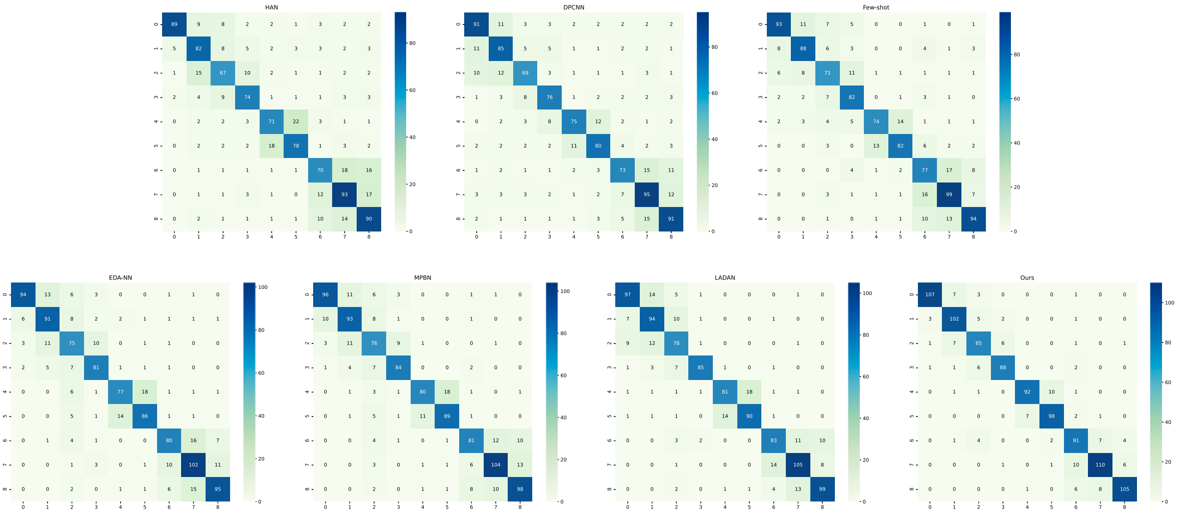


**Fig. 7.** Confusion matrix visualization of different methods on dataset Criminal-Confusing.

ensure each fold maintains the same distribution of classes as in the original dataset. We will train all models on the nine training folds and evaluate its performance on the remaining test fold. This process will be repeated ten times, with each fold used exactly once as the test set.

The results of the ten-fold cross-validation are illustrated in Fig. 6. Based on the evaluation metrics, the proposed model consistently achieves a higher performance across all metrics compared to the competing methods, indicating its superior ability in charge prediction. Furthermore, our approach maintains a better balance between precision and recall, effectively considering all charge types, thus yielding enhanced performance. In terms of model variability, the fluctuation amplitude for all methods on Criminal-Confusing is greater than that on Criminal-All, suggesting that confusing charges pose additional challenges. Across all models, there is a noticeable trend where the accuracy and Macro-F1 scores are generally higher than the Macro-recall and Micro-F1 scores. This trend indicate that while the models are generally accurate, there is room for improvement in balancing the recall across classes (as indicated by Macro-recall) and in the precision–recall balance for individual classes (as indicated by Micro-F1). We also calculated the standard deviations of the ten-fold cross-validation results for different models; the standard deviations of various metrics for the proposed model are 0.016 smaller than baselines, further confirming the increased stability of our approach.

### 5.6. Visualization

To provide an intuitive illustrate of the methodology's efficacy regarding confusing charges, and to offer a comprehensive portrayal of the proposed model's performance, we visualized the confusion matrices. In particular, we procured a random assortment of 1000 specimens from the Criminal-Confusing dataset, encompassing a total of nine typical confusing charges.[4] The visualization is shown in Fig. 7. Drawing upon these matrices, we can deduce the following:

1. Our proposed model exhibits superior overall performance, as evidenced by the elevated diagonal values signifying accurate confusing charge predictions.
2. Concerning class-specific performance, certain classes outperform others within all matrices. For instance, the first class (Intentional homicide crime) consistently boasts elevated accuracy throughout all matrices. Conversely, the third class (Picking quarrels and provoking trouble crime) possesses diminished accuracy.
3. Pertaining to the issue examined in this paper, particular classes are habitually misidentified across all matrices. For example, class 5 (Embezzlement of duty crime) is frequently misclassified as class 4 (Embezzlement crime) and vice versa. This pattern intimates that these two classes share analogous features, thus complicating the model's capacity for differentiation.
4. Some off-diagonal values remain consistently high across matrices, suggesting that specific misclassifications occur more frequently. For example, the misclassification of class 7 (Robbery crime) as class 8 (Theft crime) and class 9 (Fraud crime) constitutes a prevalent issue. This pattern denotes that these classes are more arduous to differentiate or that the model grapples with their distinction.

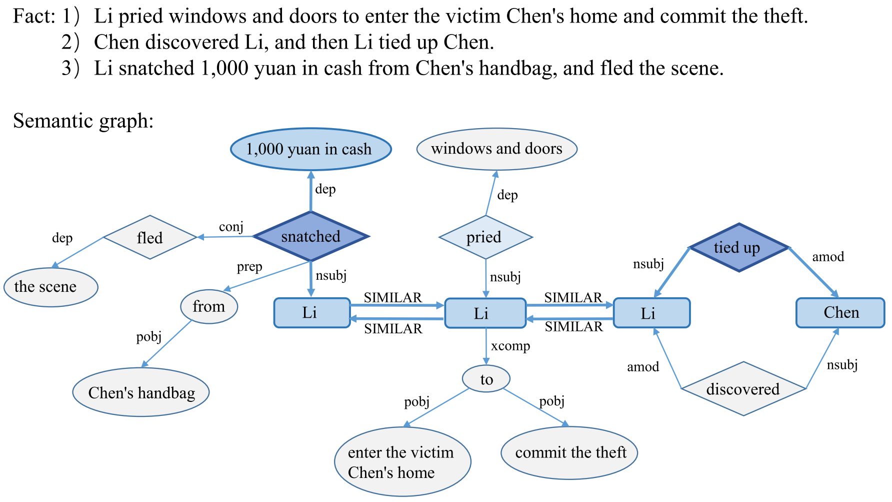
### 5.7. Analysis of semantic graph

As previously mentioned, semantic relations between entities in fact descriptions can facilitate the capture of charge-worthy content and identification of the charge type. To evaluate the efficiency of the semantic graph in selecting pertinent content, we visualize the semantic graph attention distribution for a given example. Fig. 8 presents a semantic graph constructed from a provided criminal fact. We compute the alignment of node attention using Eq. (7), with darker-colored nodes receiving more attention. From Fig. 8, it is evident that numerous significant nodes, such as "snatched" and "tied up", are captured by the proposed model. Ideally, an effective model should focus on relevant nodes while disregarding irrelevant ones, which aligns with our model's behavior. Nodes "snatched" and "tied up" represent the defendant's actions, which are crucial for predicting the charge. Consequently, the semantic graph can concentrate on pertinent content that aids in charge prediction.

### 5.8. Ablation study

Our approach is distinguished by its integration of dual-graph interaction and legal schematic knowledge. In this section, we conduct an ablation study to assess the effectiveness of these components. Firstly, we construct a uni-graph model, *w/o semantic*, by eliminating

[4] Due to space limitations, we present the predicted results of nine typical confusing charges and visualize the corresponding confusion matrix. The correspondence of these nine charges and indexes are: 0: Intentional homicide crime, 1: Intentional injury crime, 2: Picking quarrels and provoking trouble crime, 3: Intentional destruction of property crime, 4: Embezzlement crime, 5: Embezzlement of duty crime, 6: Robbery crime, 7: Theft crime, 8: Fraud crime.



Fact: 1）Li pried windows and doors to enter the victim Chen's home and commit the theft.
2）Chen discovered Li, and then Li tied up Chen.
3）Li snatched 1,000 yuan in cash from Chen's handbag, and fled the scene.

**Fig. 8.** A criminal fact example of average attention distribution on the semantic graph, with nodes colored darker for more attention.

**Table 4**
Ablation test results of two datasets.

| Datasets | Criminal-All | | | | Criminal-Confusing | | | |
|---|---|---|---|---|---|---|---|---|
| Metrics | Acc. | MR | Ma-F1 | Mi-F1 | Acc. | MR | Ma-F1 | Mi-F1 |
| *w/o semantic* | 0.912 | 0.884 | 0.899 | 0.827 | 0.891 | 0.835 | 0.849 | 0.811 |
| *w/o structure* | 0.914 | 0.887 | 0.904 | 0.834 | 0.897 | 0.839 | 0.854 | 0.819 |
| *w/o interaction* | 0.918 | 0.893 | 0.910 | 0.844 | 0.905 | 0.847 | 0.861 | 0.827 |
| *w/o knowledge* | 0.902 | 0.843 | 0.862 | 0.799 | 0.864 | 0.792 | 0.807 | 0.778 |
| **Ours** | **0.924** | **0.901** | **0.914** | **0.857** | **0.910** | **0.878** | **0.889** | **0.824** |

the semantic graph. Similarly, we create another uni-graph model, *w/o structure*, by removing the structure graph. In both models, the remaining graph is updated through the corresponding propagation rule described in Paragraph 4.2.3. Second, we modify Eq. (13) into

$$\begin{aligned} u_i^t =& r^t_{u_j} \odot u_i^{t-1} + (1 - r^t_{u_j}) \\ & \odot tanh(W_{u_h}[h^t_{u_i}; z^t_{u_j} \odot u_i^{t-1}]),\ \text{if } u_i \in v_j, \\ v_j^t =& \sum_{u_i \in v_j} r^t_{v_i} \odot v_j^{t-1} + (1 - \sum_{u_i \in v_j} r^t_{v_i}) \\ & \odot tanh(W_{v_h}[h^t_{v_j}; \sum_{u_i \in v_j} z^t_{v_i} \odot v_j^{t-1}]) \end{aligned} \tag{21}$$

to obtain the *w/o interaction* model. That is, two graphs are updated independently without any feature sharing and information interaction. And we simply concatenate node representations in two graphs to represent the fact. Finally, we construct the *w/o knowledge* model without legal schematic knowledge.

As demonstrated in Table 4, a noticeable performance decline occurs when removing any component from the two datasets. We will examine the results of the Criminal-All dataset in detail.

**Impact of semantic graph.** When the semantic graph is not utilized (*w/o semantic*), the accuracy (Acc.) score of our model decreases by 1.2%. This indicates the necessity of constructing a semantic graph to model semantic relations between relevant content for charge prediction.

**Impact of structure graph.** When the structure graph is not employed (*w/o structure*), the Acc. and Macro-F1 (Ma-F1) scores demonstrate that charge prediction based solely on the semantic graph is unsatisfactory. Consequently, the semantic graph alone is inadequate for conveying the entire meaning of the fact description. A combination with the structure graph is required for comprehensive representation.

**Impact of dual-graph interaction.** By disabling the dual-graph interaction module (*w/o interaction*), the performance drops to 0.910 in Ma-F1 compared to our full model. This suggests that sharing information through the proposed dual-graph interaction scenario is more effective than updating the structure graph and semantic graph independently. Furthermore, *w/o interaction* outperforms the two uni-graph models, highlighting the importance of employing both structure and semantic graphs simultaneously.

**Impact of legal schematic knowledge.** In the absence of legal schematic knowledge (*w/o knowledge*), the Ma-F1 score decreases from 0.857 to 0.799, illustrating the contribution of legal schematic knowledge. For a more in-depth analysis of domain knowledge, refer to Section 5.9.

In conclusion, these modules play indispensable roles in our model. For fact representation, the semantic graph captures hidden semantic relations, while the structure graph models rich global structural information. Dual-graph interaction enables the two graphs to share information with each other, resulting in a more comprehensive fact representation. Concurrently, legal schematic knowledge provides discriminative attributes about charges, and combining these modules yields improved results in the charge prediction task.

*Analysis of variance.* Given the inherent complexity of the models under review, our objective is to rigorously validate the differences in performance metrics among these models. For this purpose, we apply Analysis of Variance (ANOVA) (St, Wold, et al., 1989) to assess if the observed variances are statistically significant or merely results of random fluctuations. This method adds a layer of statistical rigor to our experimental findings. Notably, we conducted separate ANOVAs for the results of the ablation study, with these outcomes systematically detailed in the Table 6. Following this, we will succinctly explain the metrics introduced in these tables. The *sq* represents the total squared deviations linked to each factor, while the *F* metric stands for the F-statistic, which compares the variance between groups against the variance within groups. Generally, a higher F-Statistic suggests a greater disparity among the group means. Prior to the variance analysis, the two test datasets (Criminal-All and Criminal-Confusing) underwent random segmentation. Due to variations in the number of test samples between these datasets, they were divided into 300 and 100 groups, respectively. We then calculated the average values of four metrics across these segments. Thus, the degrees of freedom within groups were 1495 for Criminal-All and 495 for Criminal-Confusing, with the between-group degrees of freedom set at 4 (see Table 4), as outlined in the referenced statistical analysis table. The *P* stands for the significance level, specifically the *p*-value. A lower *p*-value indicates a stronger statistical significance of the model's data interpretation. This value is pivotal in determining the statistical significance of the F-statistic,

**Table 5**
Charge prediction results of the example case.

| Models | Gold | HAN | DPCNN | Few-shot | EDA-NN | MPBN | LADAN | Ours |
|---|---|---|---|---|---|---|---|---|
| case | *robbery* | kidnapping | kidnapping | kidnapping | kidnapping | kidnapping | kidnapping | *robbery* |

**Table 6**
The analysis of variance test for ablation study on Criminal-All and Criminal-Confusing.

| Criminal-All | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | | | MR | | | Ma-F1 | | | Mi-F1 | | |
| | *sq* | *F* | *P* | *sq* | *F* | *P* | *sq* | *F* | *P* | *sq* | *F* | *P* |
| $\mathcal{C}$ | .1169 | 218.9 | .000 | .6991 | 312.5 | .000 | .5027 | 157.2 | .000 | .5796 | 232.9 | .000 |
| $\mathcal{R}$ | .1996 | – | – | .8362 | – | – | 1.1936 | – | – | .9303 | – | – |
| $\mathcal{E}$ | .3694 | | | .4554 | | | .2961 | | | .3839 | | |
| Criminal-Confusing | | | | | | | | | | | | |
| $\mathcal{C}$ | .3773 | 534.6 | .000 | .4005 | 139.5 | .000 | .3511 | 122.5 | .000 | .1799 | 100.1 | .000 |
| $\mathcal{R}$ | .2638 | – | – | .3552 | – | – | .3546 | – | – | .2224 | – | – |
| $\mathcal{E}$ | .5885 | | | .53 | | | .4975 | | | .4472 | | |

essentially evaluating the possibility of rejecting the null hypothesis that suggests no significant difference between group means. A $p$-value below 0.05 is considered significant enough to reject the null hypothesis, demonstrating a meaningful difference in group means. Symbols $\mathcal{C}$, $\mathcal{R}$, and $\mathcal{E}$ denote the study's categorization, the unexplained variance within the models, and the effect size of an ANOVA respectively. A smaller $\mathcal{R}$ value implies a better model fit to the data, while a larger $\mathcal{E}$ value indicates a model's superior capacity to account for the observed data.

The ANOVA results provide robust statistical evidence that the performance differences across models in the ablation study are significant and not random. The analysis further reveals that these differences are more pronounced in the Criminal-Confusing dataset, highlighting the importance of the proposed modules in confusing charge prediction. Specifically,

1. In the Criminal-All dataset, the effect size values range from 0.2961 (Ma-F1) to 0.4554 (MR), suggesting that the models' differences have a moderate to strong effect on the performance metrics. This implies that the changes or modifications between the models have a considerable impact on their performance. The highest effect size is observed in MR, indicating that the model differences are most pronounced in this metric. This suggests that the models' ability to minimize false negatives varies significantly, which is critical in criminal justice where missing a true positive can have serious implications.
2. Compared to the Criminal-All dataset, the effect sizes are significantly higher across all metrics in the Criminal-Confusing dataset, with values ranging from 0.4472 (Mi-F1) to 0.5885 (Acc.). This suggests that the models' performance disparities are even more pronounced when dealing with more challenging or ambiguous cases.

### *5.9. Case study*

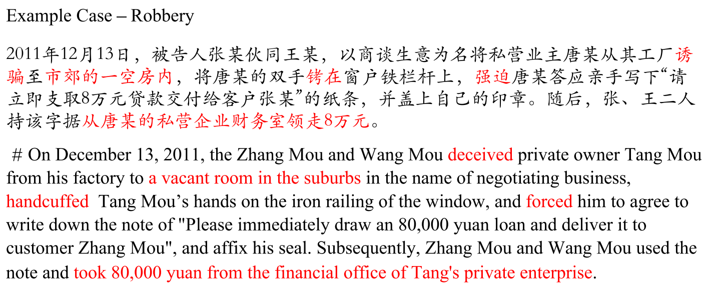
As depicted in Fig. 9, we select a representative case to provide an intuitive illustration of how legal schematic knowledge enhances the performance of confusing charge prediction. In this case, the defendant is convicted of robbery. Distinguishing between *robbery* and *kidnapping* is often challenging, as both charges involve violence and illegal possession. According to constitutive requirements, a crucial difference between them is that, in *robbery*, the defendant intends to steal property directly from the victim, whereas, in *kidnapping*, the defendant can only demand property from a third party other than the victim.

Thus, constitutive requirements play a vital role in confusing charge prediction. As seen in Table 5, only our model accurately predicts the charge as *robbery*, whereas other models incorrectly classify it as *kidnapping*. Using the Few-shot model as an example, although this case includes several attributes – such as profit motive, violence, public place, and illegal possession – these attributes are insufficiently detailed to differentiate the charges. Nevertheless, our model, leveraging legal schematic knowledge, captures the key patterns and semantics relevant to constitutive requirements, which are marked in red in Fig. 4. The prediction charge is then determined by these key patterns and semantics. This further illustrates our model's proficiency in handling confusing charges through legal schematic knowledge.

## 6. Limitations

Although our proposed method surpasses existing models to some extent, there are still several significant limitations. We will discuss these limitations from the following perspectives:

1. Latent biases: The LCP model is trained on real-world datasets, and despite our efforts to remove personal information, biases may have been unintentionally incorporated into the model, leading to inaccurate predictions. This makes it difficult to determine the root cause of the model's performance variations. The absence of information on a defendant's past experiences may cause the model to make erroneous predictions, such as for recidivism. When analyzing the experimental results, we may find that the model provides comparable predictions for similar cases, but the true labels differ significantly. Furthermore, since the datasets used consist of adjudicated texts, they do not contain any not-guilty samples. This means that the model will always render a guilty verdict.
2. Interpretability and explainability: Deep learning models, including the proposed LCP model, are often considered "black boxes" due to the difficulty in understanding and explaining their decision-making processes. Legal judgments carry high stakes, and the lack of transparency in these approaches may hinder their acceptance in the legal domain.
3. Limited applicability: The current LCP model's approach is unable to cover more complex criminal cases, such as those involving multiple defendants, which remain too challenging for our model. As a result, the proposed method has limited applicability in certain situations or for specific types of cases.
4. Availability of real data: In real legal contexts, judges review materials from various stakeholders, including public security, the prosecution service, legal representatives, and the parties involved. These materials include interrogation transcripts, party identification, investigation reports, and legal documentation, providing a multifaceted portrayal of cases that helps in the comprehensive reconstruction of factual realities. In judicial



Example Case – Robbery

2011年12月13日，被告人张某伙同王某，以商谈生意为名将私营业主唐某从其工厂诱骗至市郊的一空房内，将唐某的双手铐在窗户铁栏杆上，强迫唐某答应亲手写下“请立即支取8万元贷款交付给客户张某”的纸条，并盖上自己的印章。随后，张、王二人持该字据从唐某的私营企业财务室领走8万元。

# On December 13, 2011, the Zhang Mou and Wang Mou deceived private owner Tang Mou from his factory to a vacant room in the suburbs in the name of negotiating business, handcuffed Tang Mou’s hands on the iron railing of the window, and forced him to agree to write down the note of "Please immediately draw an 80,000 yuan loan and deliver it to customer Zhang Mou", and affix his seal. Subsequently, Zhang Mou and Wang Mou used the note and took 80,000 yuan from the financial office of Tang's private enterprise.

**Fig. 9.** An illustration of the example case, whose golden charge is **robbery**.

decision-making, the multifarious nature of sources, including audio, video, imagery, and textual data, constitutes crucial evidence. Judges meticulously validate these datasets for authenticity, legality, and pertinence, thoroughly documenting their factual and legal determinations. However, access to these datasets is often limited due to confidentiality and data protection concerns inherent in judicial proceedings. We aim to bridge this gap in the future by refining our methodology to align more closely with practical applications.

## 7. Ethical implication

As with any technological advancement, the development of legal charge prediction (LCP) models utilizing natural language processing (NLP) and deep learning approaches raises ethical concerns that warrant careful consideration. In this section, we address the ethical implications of this research with regard to data sources, data security, data privacy, and potential effects on judicial independence.

1. Data Sources: The research depends on real-world datasets to train and validate the LCP models. All data used in the experiments were publicly available. We did not engage in biased data selection, but merely transformed the format of the data. However, it is important to acknowledge that inherent biases exist within any dataset and are unavoidable.
2. Data Security: The legal domain often encompasses highly sensitive information, necessitating stringent data protection measures. In this paper, we exclusively use datasets for experimentation and do not distribute them publicly. Access to the data during the study will be granted only to authorized individuals.
3. Data Privacy: Preserving the privacy of individuals involved in legal cases is of paramount importance. To this end, we removed all content related to personally identifiable information during the experiments and utilized only a randomly assigned last name to represent a party.
4. Implications for Judicial Independence: The incorporation of LCP models into legal systems raises concerns regarding potential impacts on judicial independence. We assert that machines should not be allowed to infringe upon the independent judgment of judges at any time. Although the aim of this research is to support judges in their decision-making process and enhance efficiency, there is a risk that over-reliance on these models could compromise the human element essential to the judicial process.

## 8. Conclusion and future work

In this paper, we concentrate on the intricate task of confusing charge prediction in criminal proceedings. To the best of our comprehension, we are pioneers in employing legal domain-specific schematic knowledge as distinctive characteristics for charge prediction. We introduce a legal schematic knowledge-aware framework to address confusing charges. We construct a structural graph, a semantic graph derived from fact description, and subsequently execute dual-graph interaction to obtain fact representation, which capturing structural and semantic relationships. We incorporate the LK-Transformer to acquire better representations from legal schematic knowledge. Furthermore, we employ a knowledge-matching network to learn knowledge-aware fact representation, thereby enhancing the precision of confusing charge prediction. The empirical outcomes on real-world datasets reveal that our model attains substantial advancements over benchmarks across all evaluative criteria for criminal cases, particularly those with confusing charges. In forthcoming endeavors, it is imperative for LCP models to identify and alleviate biases within training datasets, notably implicit biases, by designing novel strategies, such as oversampling underrepresented groups or using techniques like adversarial training to reduce bias in the model’s predictions. Developing techniques to render LCP models more transparent and comprehensible enables individuals to more effectively grasp their decision-making procedures and increases trust in their predictions. In recent times, court view generation has emerged as a promising approach. With the assistance of a conversational generation pre-trained language model, the model is allowed to play the role of a judge and thus generate reasoning about the decision. Enlarging the scope of the LCP models to encompass a more extensive array of legal cases is crucial, including intricate civil and criminal cases involving multiple defendants.

## CRediT authorship contribution statement

**Sheng Bi:** Conceptualization, Methodology, Writing – original draft, Software. **Zafar Ali:** Writing – review & editing. **Tianxing Wu:** Writing – review & editing. **Guilin Qi:** Supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

The author would like to thank the reviewers for their insightful and constructive suggestions, which have been of great help in improving our manuscript. We also thank Xiya Cheng and Shiqi Zhou for writing assistance. This work supported by the Key Program of the National Social Science Foundation of China (Grant No. 21AFX013).

## References

Ahmad, S., Asghar, M. Z., Alotaibi, F. M., & Al-Otaibi, Y. D. (2022). A hybrid CNN+ BILSTM deep learning-based DSS for efficient prediction of judicial case decisions. *ESWA*, *209*, Article 118318.

Ahmed, O., et al. (2018). Artificial intelligence in HR. *IJRAR*, *5*(4), 971–978.

Aldunate, A., Maldonado, S., Vairetti, C., & Armelini, G. (2022). Understanding customer satisfaction via deep learning and natural language processing. *ESWA*, *209*, Article 118309.

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European court of human rights: A natural language processing perspective. *PeerJ Computer Science*, *2*, Article e93.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

Armour, J., & Dicker, R. (2019). Artificial intelligence in English law: A research agenda. *South Square Digest*.

Ashley, K. D., & Brüninghaus, S. (2009). Automatically classifying case texts and predicting outcomes. *AI and Law*, *17*(2), 125–165.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv:1607.06450.

Beck, D., Haffari, G., & Cohn, T. (2018). Graph-to-sequence learning using gated graph neural networks. arXiv preprint arXiv:1806.09835.

Bi, S., Ali, Z., Wang, M., Wu, T., & Qi, G. (2022). Learning heterogeneous graph embedding for Chinese legal document similarity. *KBS*, *250*, Article 109046.

Bi, S., Huang, Y., Cheng, X., Wang, M., & Qi, G. (2019). Building Chinese legal hybrid knowledge network. *Vol. 11775*, In *KSEM* (pp. 628–639).

Bruninghaus, S., & Ashley, K. D. (2003). Predicting outcomes of case based legal arguments. *AI and law*, 233–242.

Busbridge, D., Sherburn, D., Cavallo, P., & Hammerla, N. Y. (2019). Relational graph attention networks. arXiv preprint arXiv:1904.05811.

Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural legal judgment prediction in English. In *ACL* (pp. 4317–4323).

Chary, M., Parikh, S., Manini, A. F., Boyer, E. W., & Radeos, M. (2019). A review of natural language processing in medical education. *WestJEM*, *20*(1), 78.

Chen, H., Cai, D., Dai, W., Dai, Z., & Ding, Y. (2019). Charge-based prison term prediction with deep gating network. arXiv preprint arXiv:1908.11521.

Chen, S., Wang, P., Fang, W., Deng, X., & Zhang, F. (2019). Learning to predict charges for judgment with legal graph. In *ICANN* (pp. 240–252).

Dozat, T., & Manning, C. D. (2017). Deep biaffine attention for neural dependency parsing. In *ICLR*.

Gao, H., Wang, Z., & Ji, S. (2018). Large-scale learnable graph convolutional networks. In *SIGKDD* (pp. 1416–1424).

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Vol. 30*, In *NeurIPS*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

Hu, Z., Li, X., Tu, C., Liu, Z., & Sun, M. (2018). Few-shot charge prediction with discriminative legal attributes. In *COLING* (pp. 487–498).

Huang, L., Ma, D., Li, S., Zhang, X., & Wang, H. (2019). Text level graph neural network for text classification. arXiv preprint arXiv:1910.02356.

Johnson, R., & Zhang, T. (2017). Deep pyramid convolutional neural networks for text categorization. In *ACL* (pp. 562–570).

Kien, P. M., Nguyen, H.-T., Bach, N. X., Tran, V., Le Nguyen, M., & Phuong, T. M. (2020). Answering legal questions by learning neural attentive text representation. In *COLING* (pp. 988–998).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980.

Leitner, E., Rehm, G., & Moreno-Schneider, J. (2019). Fine-grained named entity recognition in legal documents. In *ICSS* (pp. 272–287).

Li, Q., Han, Z., & Wu, X.-M. (2018). Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*.

Li, S., Liu, B., Ye, L., Zhang, H., & Fang, B. (2019). Element-aware legal judgment prediction for criminal cases with confusing charges. In *ICTAI* (pp. 660–667).

Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2015). Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493.

Li, J., Zhang, G., Yu, L., & Meng, T. (2019). Research and design on cognitive computing framework for predicting judicial decisions. *JSPS*, *91*(10), 1159–1167.

Li, P., Zhao, F., Li, Y., & Zhu, Z. (2018). Law text classification using semi-supervised convolutional neural networks. In *CCDC* (pp. 309–313).

Liang, B., Su, H., Gui, L., Cambria, E., & Xu, R. (2022). Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *KBS*, *235*, Article 107643.

Lin, W.-C., Kuo, T.-T., Chang, T.-J., Yen, C.-A., Chen, C.-J., & Lin, S.-d. (2012). Exploiting machine learning models for chinese legal documents labeling, case classification, and sentencing prediction. In *ROCLING* (p. 140).

Liu, C., Chang, C., & Ho, J. (2004). Case instance generation and refinement for case-based criminal summary judgments in Chinese. *Journal of Information Science and Engineering*, *20*(4), 783–800.

Liu, Y.-H., & Chen, Y.-L. (2018). A two-phase sentiment analysis approach for judgement prediction. *JIS*, *44*(5), 594–607.

Liu, C., & Hsieh, C. (2006). Exploring phrase-based classification of judicial documents for criminal charges in Chinese. *Vol. 4203*, In *ISMIS* (pp. 681–690).

Liu, C.-L., & Liao, T.-M. (2005). Classifying criminal charges in chinese for web-based legal services. In *Asia-Pacific web conference* (pp. 64–75).

Liu, X., You, X., Zhang, X., Wu, J., & Lv, P. (2020). Tensor graph convolutional networks for text classification. *Vol. 34*, In *AAAI* (05), (pp. 8409–8416).

Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (2017). Learning to predict charges for criminal cases with legal basis. arXiv:1707.09168.

Maia, P. (2021). Intelligent compliance. In *Artificial intelligence in the economic sector* (p. 1).

Malekzadeh, M., Hajibabaee, P., Heidari, M., Zad, S., Uzuner, O., & Jones, J. H. (2021). Review of graph neural network in text classification. In *UEMCON* (pp. 0084–0091).

Marques, M. R., Bianco, T., Roodnejad, M., Baduel, T., & Berrou, C. (2019). Machine learning for explaining and ranking the most influential matters of law. In *ICAIL* (pp. 239–243).

Medvedeva, M., Vols, M., & Wieling, M. (2019). Using machine learning to predict decisions of the European court of human rights. *AI and Law*, 1–30.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR*.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *ICML* (pp. 807–814).

Ragesh, R., Sellamanickam, S., Iyer, A., Bairi, R., & Lingam, V. (2021). Hetegcn: heterogeneous graph convolutional networks for text classification. In *WSDM* (pp. 860–868).

Rosili, N. A. K., Zakaria, N. H., Hassan, R., Kasim, S., Rose, F. Z. C., & Sutikno, T. (2021). A systematic literature review of machine learning methods in predicting court decisions. *IJAI*, *10*(4), 1091.

Ruiz, L., Gama, F., & Ribeiro, A. (2020). Gated graph recurrent neural networks. *TSP*, *68*, 6303–6318.

Shaikh, R. A., Sahu, T. P., & Anand, V. (2020). Predicting outcomes of legal cases based on legal factors using classifiers. *Procedia Computer Science*, *167*, 2393–2402.

Shen, Y., Sun, J., Li, X., Zhang, L., Li, Y., & Shen, X. (2018). Legal article-aware end-to-end memory network for charge prediction. In *Proc. int. conf. eng. sci. appl.* (pp. 1–5).

St, L., Wold, S., et al. (1989). Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, *6*(4), 259–272.

Sulea, O.-M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P., & Van Genabith, J. (2017). Exploring the use of text classification in the legal domain. arXiv preprint arXiv:1710.09306.

Sulea, O.-M., Zampieri, M., Vela, M., & Van Genabith, J. (2017). Predicting the law area and decisions of french supreme court cases. arXiv preprint arXiv:1708.01681.

Surden, H. (2019). Artificial intelligence and law: An overview. *Georgia State University Law Review*, *35*, 19–22.

Undavia, S., Meyers, A., & Ortega, J. E. (2018). A comparative study of classifying legal documents with neural networks. In *FedCSIS* (pp. 515–522).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *NeurIPS* (pp. 5998–6008).

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. arXiv preprint arXiv:1710.10903.

Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., et al. (2019). Heterogeneous graph attention network. In *WWW* (pp. 2022–2032).

Wei, D., & Lin, L. (2019). An external knowledge enhanced multi-label charge prediction approach with label number learning. arXiv:1907.02205.

Wei, F., Qin, H., Ye, S., & Zhao, H. (2018). Empirical study of deep learning for text classification in legal document review. In *ICBD* (pp. 3317–3320). IEEE.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *TNNLS*, *32*(1), 4–24.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., & Weinberger, K. (2019). Simplifying graph convolutional networks. In *ICML* (pp. 6861–6871).

Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., et al. (2018). Cail2018: A large-scale legal dataset for judgment prediction. arXiv:1807.02478.

Xu, N., Wang, P., Chen, L., Pan, L., Wang, X., & Zhao, J. (2020). Distinguish confusing law articles for legal judgment prediction. arXiv:2004.02557.

Yang, W., Jia, W., Zhou, X., & Luo, Y. (2019). Legal judgment prediction via multi-perspective bi-feedback network. In *IJCAI* (pp. 4085–4091).

Yang, Z., Wang, P., Zhang, L., Shou, L., & Xu, W. (2019). A recurrent attention network for judgment prediction. In *ICANN* (pp. 253–266).

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *NAACL* (pp. 1480–1489).

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *Vol. 33*, In *AAAI* (01), (pp. 7370–7377).

Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *CSN*, *6*(1), 1–23.

Zhang, S., Zheng, D., Hu, X., & Yang, M. (2015). Bidirectional long short-term memory networks for relation classification. In *PACLIC* (pp. 73–78).

Zhong, H., Wang, Y., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). Iteratively questioning and answering for interpretable legal judgment prediction. *Vol. 34*, In *AAAI* (01), (pp. 1250–1257).

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How does NLP benefit legal system: A summary of legal artificial intelligence. arXiv preprint arXiv:2004.12158.