Expert Systems With Applications 261 (2025) 125462

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Simulating judicial trial logic: Dual residual cross-attention learning for predicting legal judgment in long documents

Congqing He [a], Tien-Ping Tan [a,*], Sheng Xue [b], Yanyu Tan [c,*]

[a] School of Computer Sciences, Universiti Sains Malaysia, Penang, 11800, Malaysia
[b] Guangdong Research Institute, China Telecom Corporation Ltd., Guangzhou, 510006, China
[c] School of Public Administration and Human Geography, Hunan University of Technology and Business, Changsha, 410205, China

ARTICLE INFO

Dataset link: https://data.thunlp.org/legal/CAIL-Long.tar.gz

Keywords:
Legal judgment prediction
Lengthy legal documents
Judicial trial logic
Dual residual cross-attention
Constrained cross-entropy loss

ABSTRACT

Legal Judgment Prediction (LJP) plays a vital role in judicial assistance systems, aiming to predict judgment outcomes automatically from the fact descriptions in legal cases. A key challenge in LJP lies in effectively capturing decisive information that influences legal judgments — such as criminal events, behaviors, and consequences — especially from lengthy legal documents. Existing approaches, which incorporate domain-specific knowledge, have attempted to improve the ability to capture decisive information but often fail to adequately address the complexities of longer texts and may introduce noise, leading to incorrect predictions. To overcome these limitations, we propose a novel method, **JuriSim**, for predicting Chinese criminal legal judgments in long documents by incorporating the knowledge of judicial trial logic. Specifically, JuriSim extracts legal events and generates rationales based on fact descriptions to capture the decisive information that influences judgment outcomes. Then, a dual residual cross-attention mechanism is introduced to interactively process facts, events, and rationales for predicting relevant legal statutes, charges, and term of penalties. This mechanism allows the model to reduce the loss of important information and the retention of incorrect information during the aggregation process. Furthermore, we present a constrained cross-entropy loss, utilizing the topological relationship between charges, terms, and applicable law statutes. Experiments conducted on the publicly available CAIL-Long criminal dataset demonstrate the efficiency of the JuriSim framework in predicting legal judgments, especially for cases involving long documents. JuriSim_Lawformer shows a relative improvement of 3.45% in Macro-F1 for charge prediction and 3.05% for term of penalty prediction, compared to Lawformer.

1. Introduction

Artificial Intelligence (AI) in the legal domain has been studied for decades (Hu, Li, Tu, Liu, & Sun, 2018; Le, Xiao, Xiao, & Li, 2023; Lyu et al., 2022; Makridakis, 2017). Legal Judgment Prediction (LJP) is a process to predict the judgment outcomes from the fact description of legal cases. In real-world scenarios, LJP can automatically push case analyses, legal statutes, and judgment outcomes during the judges' case handling processes, providing them with unified and comprehensive trial standards and case handling guidelines. Based on predictive judgments, the system automatically alerts to any significant deviations between the judgments produced by judges and the predicted outcomes, thereby preventing major deviations in judicial standards. On the other hand, by publicly sharing and analyzing LJP's results, it can enhance public understanding and trust in the criminal trial process, increasing the transparency of the legal system and public acceptance. This paper addresses LJP in Chinese criminal legal cases, focusing on predicting three sub-tasks: the applicable law statutes (Li, Ge, Cheng, Luo, & Chang, 2022), the charges (Hu et al., 2018), and the penalty terms (Chen, Cai, Dai, Dai, & Ding, 2019).

Despite the success of deep learning and pre-trained language models in the legal arena (Luo, Feng, Xu, Zhang, & Zhao, 2017; Mamakas, Tsotsi, Androutsopoulos, & Chalkidis, 2022), LJP has encountered several challenges. One of the main challenges is the lengthy fact descriptions in legal cases. For example, we observed that 19% of the fact descriptions in the CAIL-Long criminal dataset exceed 1000 tokens in length, as depicted in Fig. 1(a). Lengthy legal documents complicate the capture of decisive information that influences legal judgments. Crucial information, such as key events and criminal behaviors, may be intermixed with a vast amount of irrelevant information. For instance, Fig. 1(c) presents a performance comparison across different fact description length groups in legal judgment prediction tasks. The

* Corresponding authors.
E-mail addresses: hecongqing@student.usm.my (C. He), tienping@usm.my (T. Tan), xues1@chinatelecom.cn (S. Xue), ytan1949@hutb.edu.cn (Y. Tan).

https://doi.org/10.1016/j.eswa.2024.125462
Received 3 May 2024; Received in revised form 9 September 2024; Accepted 24 September 2024
Available online 29 September 2024

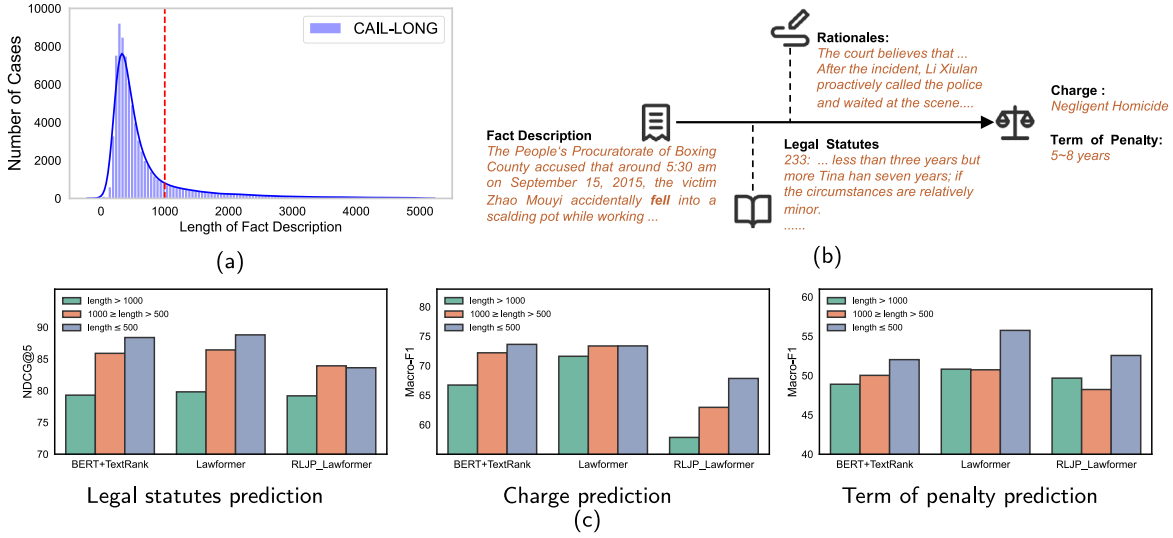

Fig. 1. (a) Distribution of fact description lengths in the CAIL-Long criminal datasets. (b) The decision-making process follows the logic of judicial trials. (c) Performance comparison of different fact description length groups in legal judgment prediction tasks.

performance is noticeably better for tasks with fact lengths less than 500 tokens but significantly worsens for lengths greater than 1000 tokens. This decline is primarily because key events and criminal behaviors may be obscured by a vast amount of irrelevant information, adversely affecting the performance of LJP.

To mitigate this issue, existing approaches has incorporated domain-specific knowledge into LJP systems, focusing on extracting relevant events and rationales from extensive texts to bolster model efficacy (Feng, Li, & Ng, 2022; Wu et al., 2022; Xu et al., 2020; Yue, Liu, Wu et al., 2021). For example, Feng et al. (2022) proposed a model that enhances judgment prediction by extracting and utilizing legal events, enforced by cross-task consistency constraints. Wu et al. (2022) integrated rationales to bolster both the accuracy and interpretability of predictions. However, these approaches often rely on single-dimensional domain-specific knowledge and fail to adequately address the complexities of longer texts. Moreover, these methods overlook the noise present in the extracted domain-specific knowledge and merge this knowledge as inputs, which may lead to incorrect judgment predictions.

In this paper, we adopt the knowledge of judicial trial logic to predict legal judgments. Fig. 1(b) presents the decision-making process according to the logic of judicial trials. The knowledge of judicial trial logic involves multiple parts. Firstly, it involves an analysis of fact descriptions. Subsequently, judges identify applicable legal statutes. The next step involves summarizing criminal behaviors and consequences from the fact description to formulate rationales. The final stage is the determination of judgment outcomes (including charges and term of penalties). This knowledge plays a crucial role in LJP:

- **Events:** Events are key information extracted from fact descriptions, including key criminal events.
- **Rationales:** Rationales are summaries based on fact descriptions, encapsulating criminal behaviors and consequences that influence the judgment outcome.
- **Legal Statutes:** Legal statutes are fundamental in countries using the civil law system, such as China (excluding Hong Kong), and are essential to judicial decision-making. For instance, Statute 245 states: *"Individuals who conduct illegal searches of another person's body or residence, or unlawfully enter someone else's residence, shall be punished with up to three years of fixed-term imprisonment or criminal detention. Judicial personnel who abuse their authority and commit the aforementioned crimes shall be subject to more severe penalties"*. This statute involves charges such as *"Illegal Search"*, *"Illegal Home Invasion"*, and *"Judicial Staff Abuse of Power"*, with potential penalty terms ranging from 0 to 3 years.

Therefore, to determine the judgment outcome for a case, we first analyze the key information and summarize vital details related to criminal behaviors and consequences. Simultaneously, we identify the relevant legal statutes involved in the case.

To this end, we propose a novel framework named JuriSim for predicting legal judgments by integrating the knowledge of judicial trial logic, which comprises four stages. In the first stage, the framework extracts and recognizes events and event types from the fact descriptions, aiming to analyze key events in the legal case. This is followed by the use of interactive representations of both the fact descriptions and event sequences to predict applicable legal statutes. Simultaneously, the framework generates rationales based on the fact descriptions, aiming to capture criminal behaviors and consequences in the legal case. The final stage involves the interactive processing of facts, events, and rationales to predict judgment outcomes, including charges and term of penalties. In addition, the determination of charges and penalty terms correlates with applicable legal statutes. Therefore, we introduce a constrained cross-entropy loss that leverages the topological relationship between sub-tasks. Each legal statute is associated with a subset of charges and terms, and these topological relationships are integrated into the constrained loss function.

To address the issue of noise in the extracted domain-specific knowledge, which may lead to incorrect judgments prediction, we propose a dual residual cross-attention mechanism that interactively processes facts, events, and rationales. First, a residual cross-attention mechanism integrates fact descriptions and event sequences, allowing the model to focus on key events while reducing both the loss of important information and the retention of incorrect information during the aggregation process. Subsequently, we aggregate rationales and fact descriptions using another residual cross-attention mechanism, effectively capturing vital information related to criminal behaviors and consequences, and minimizing the loss of important information and the retention of incorrect data.

We have summarized the main contribution as below:

(1) We present a novel framework for predicting Chinese criminal legal judgments by integrating the knowledge of judicial trial logic. This framework incorporates events to predict applicable legal statutes, and then integrates events and rationales to predict charges and penalty terms. This approach enhances the model's ability to capture decisive information that influences judgment predictions, such as criminal events, behaviors, and consequences.

(2) We introduce a dual residual cross-attention mechanism that interactively processes facts, events, and rationales. This mechanism

allows the model to reduce the loss of important information and the retention of incorrect information during the aggregation process.

(3) We introduce a constrained cross-entropy loss, leveraging the topological relationship between applicable legal statutes and charges, term of penalties.

(4) Experiments conducted on a publicly available CAIL-Long criminal dataset demonstrate the efficiency of the JuriSim framework in predicting legal judgments, especially for long documents cases.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 describes our proposed method in detail. Sections 4 and 5 provide the experimental setup, results, and analysis. Section 6 concludes the paper and discusses future work. Finally, Section 7 and Section 8 present the ethical discussion and limitations of this paper.

2. Related work

2.1. Legal judgment prediction

LJP has been researched over several decades, and various legal systems, such as China (Le et al., 2023; Luo et al., 2017), the US (Katz, Bommarito, & Blackman, 2017), Europe (Chalkidis, Androutsopoulos, & Aletras, 2019), French (Şulea, Zampieri, Vela, & van Genabith, 2017), Switzerland (Alali, Syed, Alsayed, Patel, & Bodala, 2021), Indian (Malik et al., 2021), Korean (Hwang, Lee, Cho, Lee, & Seo, 2022), Arabic (Almuzaini & Azmi, 2023). In the earlier years, LJP models employed rule-based and machine learning approaches, necessitating manual feature extraction (Katz et al., 2017; Kort, 1957; Şulea et al., 2017). Although these methods were straightforward and reliable, there are labor intensive process of feature extraction. With recent advancements in AI, an increasing number of researchers explored the utilization of deep neural networks to address the LJP problem. This section presents an overview of the recent literature on LJP, which focuses on two problems: incorporating of judicial trial logic, and handling long legal documents.

A branch of LJP-related work focuses on incorporating domain-specific knowledge to improve performance, such as the relevant legal statutes, event extraction, rationales, similar cases, and the dependencies between the tasks of LJP. Luo et al. (2017) devised a method using an attention-based deep neural network. This approach aims to predict criminal charges based on fact descriptions, coupled with the extraction of relevant legal statutes. Xu et al. (2020) employed an attention-based graph neural network, to distinguish between similar legal statutes and to extract distinctive representation from fact descriptions. Yue, Liu, Jin et al. (2021) extracted crime-related snippets from the fact description and used the snippets to predict judgment decisions. Feng et al. (2022) introduced a constrained, event-based judgment prediction model. This model leverages the extraction of events from case facts and employs cross-task consistency constraints to enhance performance. Wu et al. (2022) introduced rationale to enhance both the performance and interpretability of LJP. The method splits the prediction process into two steps: generating rationales based on the fact, and predicting the judgment based on the generated rationales. Liu, Du, Li, Pan, and Ming (2022) proposed a LJP framework with case triple modeling, which samples similar and dissimilar cases to construct a case triple based on contrastive case relations. The model then refines the encoding and decoding processes to obtain the predicted labels for each LJP-based sub-task. Despite these advancements, these studies often rely on single-dimensional domain-specific knowledge and fail to adequately address the complexities of longer texts. For instance, Wu et al. (2022) did not consider that the process of rationale generation can yield incorrect information, and the attention mechanisms integrating facts with noisy rationales might mislead the model, leading to inaccurate judgment predictions.

Several studies have focused on the dependencies between the tasks of charges, terms, and applicable law statutes, utilizing multi-task learning for LJP. For instance, Zhong et al. (2018) introduced a multi-task learning method for LJP, employing scalable Directed Acyclic Graph structures to explicitly model the dependencies among legal statutes, charges, and terms of penalties. Yang, Jia, Zhou, and Luo (2019) explored the interrelations within LJP outcomes related to law statutes, charges, and term of penalties. The method introduced a word collocation attention mechanism that extracts information on word collocations and the semantics of numbers from fact descriptions for judgment predictions. However, while these methods leverage task dependencies to enhance the effectiveness of LJP, inaccuracies in sub-task predictions can result in errors propagating to other tasks.

LLMs have achieved strong performance in a wide range of NLP tasks through prompt-based learning (Liu et al., 2023). Several studies have explored the use of LLMs as foundational models and leveraged prompt learning to enhance the effectiveness and explainability of LJP (Deng, Mao, Zhang, & Dou, 2024; Jiang & Yang, 2023; Wu et al., 2023). For example, Wu et al. (2023) introduced the Precedent-Enhanced Legal Judgment Prediction framework, which utilizes historical precedents by combining LLMs with domain-specific models to predict legal judgments. The domain models provide candidate labels and retrieve relevant precedents, while LLMs make the final predictions by understanding the context of these precedents. Jiang and Yang (2023) introduced the Legal Syllogism Prompting (LoT) method, designed to guide large language models in LJP without the need for additional learning or fine-tuning. This approach integrates the structure of legal syllogism into the model's prompting process, utilizing legal statutes as the major premise, case facts as the minor premise, and court judgments as the conclusion. Deng et al. (2024) introduced the ADAPT (Ask-Discriminate-Predict) framework for LJP, which mimics human judicial reasoning by decomposing case descriptions, distinguishing potential charges, and predicting final judgments. These methods demonstrate competitive performance in the interpretability of legal judgments, but zero-shot-based LLMs show limited performance in LJP (Wu et al., 2023).

Another branch of related work focuses on addressing long text problems in the legal domain. Wan, Papageorgiou, Seddon, and Bernardoni (2019) introduced a hierarchical network for processing lengthy legal texts. This approach involves segmenting the text and then processing the output representations of these segments using a BiLSTM to yield a unified document representation. Lawformer (Xiao, Hu, Liu, Tu, & Sun, 2021) is used for processing long Chinese legal documents. It was pre-trained on a large corpus of extensive Chinese legal cases, employing a masked language modeling objective and a Longformer-based architecture (Beltagy, Peters, & Cohan, 2020). Mamakas et al. (2022) explored two approaches for handling long legal texts in English: adapting a Longformer model, pre-trained on LegalBERT, to handle longer legal texts, and modifying LegalBERT to utilize TF-IDF representations. However, these works mainly rely on PLMs to represent long documents without fully utilizing the key information and events of criminal behavior for LJP.

2.2. Long document classification

The advent of PLMs (Kenton & Toutanova, 2019) has accentuated the issue of document length. For instance, BERT processes input sequences of up to 512 tokens, and the common approach for handling longer texts is to truncate them to 512 tokens. However, this truncation method omits crucial information, potentially leading to reduced model performance. To tackle the challenges with long document classification, two main approaches have been proposed: Hierarchical Transformers and Resource-Efficient Transformers.

Hierarchical Transformers segment long documents into shorter, manageable lengths and then process them independently to produce corresponding segment-level semantic representations. Chalkidis et al. (2019) developed a hierarchical BERT based method to overcome length limitations for LJP in English. This method demonstrated superior performance compared to both standard BERT and the original HANs (Yang et al., 2016). ToBERT (Pappagari, Zelasko, Villalba,

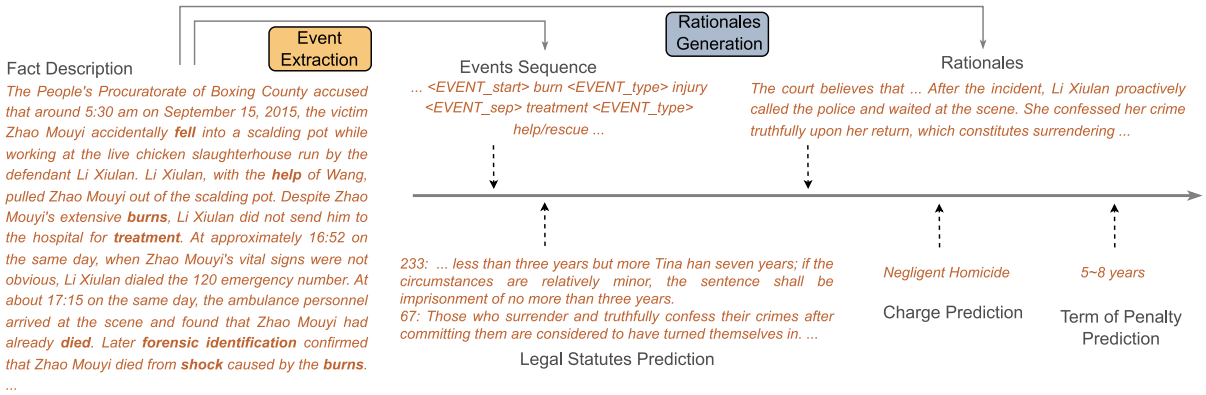

Fig. 2. The flow chart of the JuriSim framework. The process first involves extracting events from a fact description, integrating them to identify relevant legal statutes, and concurrently generating rationales. Finally, the framework predicts charges and term of penalties.

Carmiel, & Dehak, 2019) is a method that enhances BERT's ability to classify long documents by applying an additional Transformer layer on top of segmented BERT representations. This approach overcomes BERT's limitations in handling long sequences and reduces computational complexity.

Transformer-based PLMs require a large number of computational resources to process long texts, due to the time complexity of the self-attention mechanism. Resource-Efficient Transformers have been introduced to alleviate the complexity challenges of self-attention mechanisms, benefiting the efficient processing of long texts. For example, Longformer (Xiao et al., 2021) introduced an attention mechanism with a time complexity that is linearly related to the sequence length, enabling it to handle documents containing thousands of tokens. Inspired by graph sparsification, Zaheer et al. (2020) introduced BigBird, which employs a sparse attention mechanism. The complexity of this mechanism is linearly related to the sequence length (i.e., it increases linearly with the length of the text), enhancing the model's performance on long-sequence texts. Furthermore, Large Language Models (LLMs), such as LLAMA2 (Touvron et al., 2023) and Qwen-2 (Yang et al., 2024), also employ similar methods to handle longer texts. For instance, LLAMA2 adopts Grouped-Query Attention (Ainslie et al., 2023) to process longer text.

To evaluate the comparative performance of Hierarchical Transformers and Resource-Efficient Transformers, Dai, Chalkidis, Darkner, and Elliott (2022) conducted experiments using various Transformer-based models, including Longformer and hierarchical transformers, for long document classification. Their study demonstrates that both Longformer and hierarchical transformers outperform BERT-based models.

2.3. Chinese natural language processing in legal texts

The integration of NLP techniques with Chinese language processing presents unique challenges and opportunities, especially in the legal field (Wong, Li, Xu, & Zhang, 2022; Zhong et al., 2020). The distinctive linguistic features of Chinese, such as the absence of explicit word boundaries and the prevalence of homophones, require segmentation algorithms and contextual analysis to ensure precise text analysis (Huang & Powers, 2003). Additionally, the syntactic ambiguity and semantic richness of Chinese pose significant challenges for automated systems, necessitating sophisticated models that can capture nuanced meanings.

Recent years have witnessed substantial enhancements in Chinese Natural Language Processing (CNLP) technologies, particularly in semantic understanding and contextual analysis (Cui et al., 2020). Techniques such as deep neural networks and transfer learning have been adapted to better handle the complexities of Chinese syntax and semantics (Cui, Che, Liu, Qin, & Yang, 2021; Yang et al., 2024; Zhang et al., 2021). For instance, BERT-based transformer models, trained on vast corpora of Chinese texts, are adapted to process and understand Chinese grammar and vocabulary (Cui et al., 2021). Furthermore, these models, trained on extensive corpora of legal texts, significantly improve semantic understanding and contextual analysis capabilities on legal texts, dramatically impacting performance in legal tasks (Xiao et al., 2021). Specifically, the application of BERT-based models trained on legal Chinese texts has shown promising results in predicting legal judgments (Zhong, Zhang, Liu, & Sun, 2019).

3. Methodology

3.1. Model formulation

In this section, we present the JuriSim framework. Fig. 2 displays a flow chart detailing the processes within JuriSim. Specifically, given a fact description F, the framework firstly extracts events and the corresponding event types, represented as E. Subsequently, it integrates the fact description F and the event sequence E to find applicable legal statutes. Simultaneously, based on the fact description F, it extracts and generates rationales, denoted as R. Finally, considering the fact description F, event sequence E, rationales R, and the topological relationship between legal statutes, charges, and term of penalties, the framework predicts charges and term of penalties. For training convenience, event extraction and rationales generation are trained separately. Legal statutes, charges, and term of penalties are trained within a unified framework.

3.2. Events extraction

A two-stage method is adopted for the extraction of legal-oriented events and event types, following the previous work on legal event detection (Yao et al., 2022). In the first stage, potential legal-related events are extracted from the fact description using BERT-CRF (Souza, Nogueira, & Lotufo, 2019). Subsequently, DMBERT (Wang, Han, Liu, Sun, & Li, 2019) is utilized to confirm the legal relevance of the extracted events and to identify their specific event types. The representation of events is not the focus of this study. The method for the event extraction process is depicted in Fig. 3.

BERT-CRF: The BERT-CRF model integrates three components: a BERT encoder, a token-level classifier and a CRF (Lafferty, McCallum, & Pereira, 2001). Given an input sequence of l_F tokens $\{w_1, \ldots, w_{l_F}\}$, the BERT encoder first generates a sequence of hidden embeddings $\{h_1, \ldots, h_{l_F}\}$. These embeddings are then mapped to tag spaces by the token-level classifier. Finally, the classifier's outputs are inputted into the CRF layer to obtain the optimal sequence tagging.

DMBERT: DMBERT utilizes the BERT encoder with a dynamic multi-pooling operation. For an input sequence of l_F tokens $\{w_1, \ldots,$

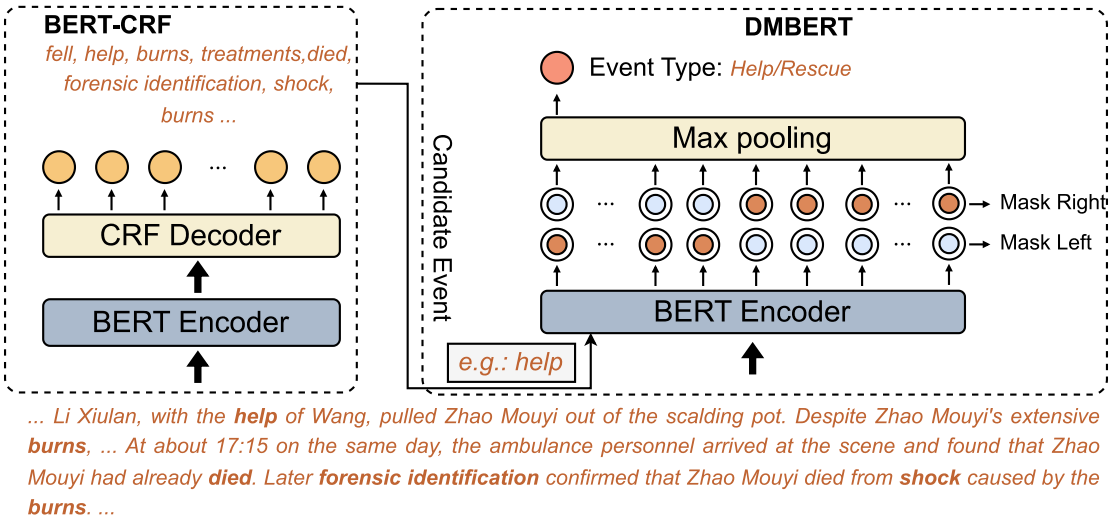

Fig. 3. The method of the event extraction process.

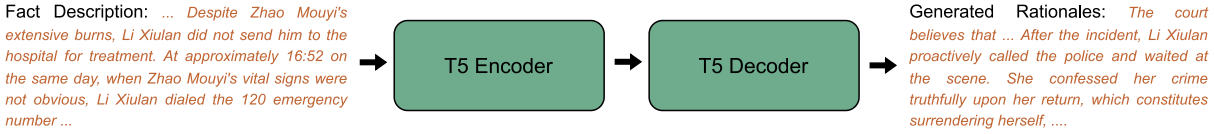

Fig. 4. The method of the rationale generation process.

$w_t, \ldots, w_{l_F}\}$, where w_t represents the candidate tokens, BERT outputs a sequence of hidden embeddings $\{h_1, \ldots, h_t, \ldots, h_{l_F}\}$. Then, these embeddings are fed into a dynamic multi-pooling to obtain an interactive representation of the candidate event and the input sequence. Specifically, the input sequence is divided by the candidate token w_t, obtaining two segments: $\{h_1, \ldots, h_t\}$ and $\{h_{t+1}, \ldots, h_{l_F}\}$. These segments are individually fed to dynamic multi-pooling and subsequently concatenated. The event type of the candidate event is then predicted by passing the concatenated output representation.

3.3. Rationales generation

Following the previous work on rationale extraction and generation (He, Tan, Xue, & Tan, 2023; Wu et al., 2022; Yue, Liu, Wu et al., 2021), we adopt the T5 architecture (Raffel et al., 2020) for generating rationales from fact descriptions. T5 uses a standard Transformer (Vaswani et al., 2017) architecture with an encoder and a decoder. As depicted in Fig. 4, the method takes the fact description as input, then processes it through the T5 model's encoder and decoder modules, end-to-end generating rationales.

3.4. JuriSim framework

The JuriSim framework for predicting legal judgments is depicted in Fig. 5. The left figure presents the overall framework, while the right figure provides a detailed view of the Residual Cross-Attention Mechanism implemented within it. Firstly, the event extraction model and rationale generation model are used to generate the events sequence and rationales, which are introduced in Sections 3.2 and 3.3. Then, PLMs are utilized to represent events E, rationales R, and fact descriptions F respectively. Subsequently, a dual residual cross-attention mechanism is introduced to interactively process the representations of facts $\mathbf{H_f}$, events $\mathbf{H_e}$, and rationales $\mathbf{H_r}$. The final predictions of legal statutes, charges, and term of penalties are based on these representations: $\mathbf{H_f}$, $\mathbf{H_e}$ and $\mathbf{H_r}$, as well as the interactions between events and facts $\mathbf{H_{fe}}$, and the interaction of rationale and facts $\mathbf{H_{fr}}$. Furthermore, our framework includes a constraint loss to consider the interrelationships among sub-tasks.

3.4.1. Events representation

Events are extracted from the fact description, which can capture key event information that determines the judgment. To represent these events, we serialize the event sequence extracted in Section 3.2 to plain text and then utilize PLMs for encoding. Specifically, we separate the events introduced in Section 3.2 with special tokens, e.g., "<EVENT_start> help <EVENT_type> help/rescue <EVENT_sep>... <EVENT_end>", where "<EVENT_start>", "<EVENT_type>", "<EVENT_sep>", "<EVENT_end>" represent the start, type, separation, and end of events, respectively. In our approach, BERT is employed to encode the event sequence, and obtains the corresponding sequences of hidden states, represented as $\mathbf{H_e}$.

$$\mathbf{H_e} = \text{Encoder}_e(E). \tag{1}$$

3.4.2. Rationales representation

Similarly, rationales are generated from the fact description, which can capture key information that determines the judgment, including criminal behavior, consequences, and circumstances. To represent these rationales, we adopt a method similar to that used in events representation. This involves encoding the rationales using BERT, yielding the corresponding sequences of hidden states, denoted as $\mathbf{H_r}$.

$$\mathbf{H_r} = \text{Encoder}_r(R). \tag{2}$$

3.4.3. Fact description representation

To investigate the performance of Hierarchical Transformers and Resource-Efficient Transformers on LJP for Chinese long legal texts, we utilize two kinds of PLMs in representing fact description.

Hierarchical Transformers: The fact description, denoted as $F = \{w_t\}_{t=1}^{l_F}$, is firstly divided into multiple segments, with each segment containing less than 512 tokens. Each segment is then independently encoded using these encoders. Following this, the segmented representations are processed through an additional Transformer layer, which is finally aggregated into a document representation.

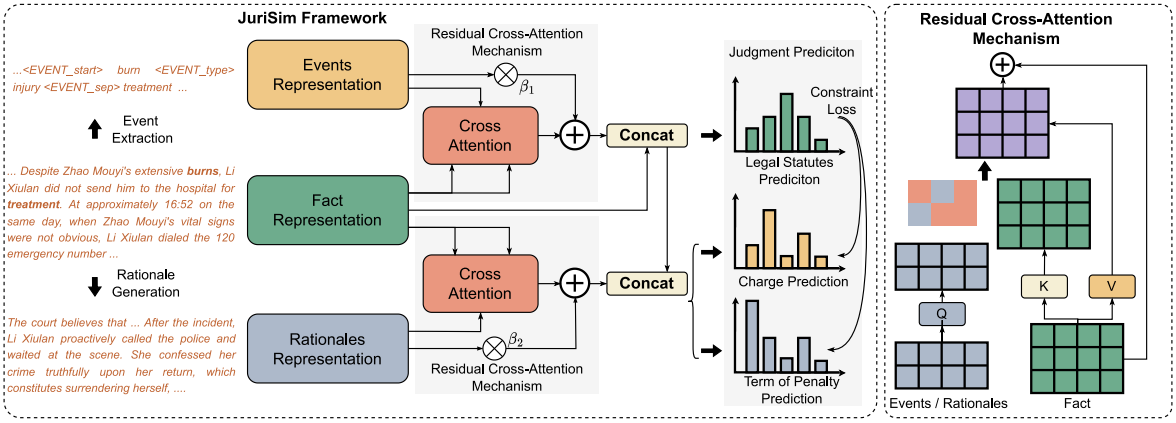

Fig. 5. The Jurisim framework for predicting legal judgments in long documents. The left figure illustrates the structure of the JuriSim framework, and the right figure shows the residual cross-attention mechanism.

Longformer/Lawformer: The fact description, represented as a sequence of tokens and denoted as $F = \{w_t\}_{t=1}^{l_F}$, is fed into PLMs, such as Lawformer for token encoding (Xiao et al., 2021).

$$\mathbf{H_f} = \text{Encoder}_f(F), \tag{3}$$

where $\mathbf{H_f}$ denotes the hidden representations of the fact description.

3.4.4. Dual residual cross-attention mechanism

A dual residual cross-attention mechanism is introduced to interactively process facts, events, and rationales within our model. This mechanism significantly enhances the model's capability to capture the key events and vital information related to criminal behaviors and consequences from the fact descriptions. Specifically, the cross-attention mechanism is employed to facilitate interaction between the fact description and the event sequence. The formulation of this mechanism is as follows:

$$Q_i = W_i^Q \mathbf{H_e}, \tag{4}$$

$$K_i = W_i^K \mathbf{H_f}, \tag{5}$$

$$V_i = W_i^V \mathbf{H_f}, \tag{6}$$

$$A_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i, \tag{7}$$

$$\mathbf{H_{fe}} = \text{Concat}(A_1, \ldots, A_m) W^M, \tag{8}$$

where $\mathbf{H_f}$ and $\mathbf{H_e}$ denote the extracted representations from F and E, respectively. The parameters W_i^Q, W_i^K, W_i^V, and W^M are trainable within the model. In the attention mechanism, for each head i, the corresponding attention A_i of the ith head is calculated as a weighted sum of the features. There are a total of m heads in the mechanism. The $\mathbf{H_{fe}}$ represents the attentions of ongoing events under the fact description.

Similarly, the cross-attention mechanism is employed for rationales and fact descriptions. The obtained $\mathbf{H_{fr}}$, represents the attentions of rationales under the fact description.

Furthermore, we introduce residual mechanism to reduces the loss of important information and the retention of incorrect information during the aggregation process.

$$\mathbf{H_{he}} = \beta_1 \cdot \mathbf{H_e} + \mathbf{H_{fe}}, \tag{9}$$

$$\mathbf{H_{hr}} = \beta_2 \cdot \mathbf{H_r} + \mathbf{H_{fr}}, \tag{10}$$

where β_1 and β_2 are the scale factors of $\mathbf{H_e}$ and $\mathbf{H_r}$, respectively, with values ranging between 0 and 1. The $\cdot$ represents the scalar multiplication of the β_1 and β_2 with the $\mathbf{H_e}$ and $\mathbf{H_r}$, respectively.

3.4.5. Judgment prediction

Given the representations of $\mathbf{H_f}$, $\mathbf{H_{he}}$, and $\mathbf{H_{hr}}$, the three sub-tasks predict the corresponding results respectively. Specifically, for the prediction of law statutes, a fully connected layer (FC) coupled with a sigmoid function is utilized to obtain the probability distribution of legal statute labels.

$$P(article) = \text{Sigmoid}(\text{FC}(\left[\mathbf{H_f}, \mathbf{H_{he}}\right])). \tag{11}$$

The model employs a fully connected layer (FC) and a softmax function to predict charges and penalty terms, obtaining the probability distribution for the labels.

$$P(charge) = \text{Softmax}(\text{FC}(\left[\mathbf{H_f}, \mathbf{H_{he}}, \mathbf{H_{hr}}\right])), \tag{12}$$

$$P(term) = \text{Softmax}(\text{FC}(\left[\mathbf{H_f}, \mathbf{H_{he}}, \mathbf{H_{hr}}\right])), \tag{13}$$

where the fully connected layer (FC) consists of two linear layers.

3.4.6. Training

Considering that each legal statute is associated with a subset of charges and terms, and given the topological dependencies among sub-tasks in LJP, we first identify applicable legal statutes from fact descriptions, and then determine the charges and terms of penalties based on these legal statutes. Therefore, our approach introduces a constrained cross-entropy loss for predicting both charges and term of penalties, utilizing the topological relationships between sub-tasks.

Suppose a cross-entropy loss, represented as L, where y_{pred} represents the predicted values and y_{true} the true values. We introduce a constraint range $[a, b]$, and a penalty coefficient λ with a value range of [0-1]. For the task of term of penalty prediction, the penalty function L_P is define as:

$$L_P(y_{\text{pred}}, \lambda) = \begin{cases} 0 & \text{if } a \le y_{\text{pred}} \le b, \\ \lambda \cdot \min(|y_{\text{pred}} - a|, |y_{\text{pred}} - b|) & \text{otherwise}, \end{cases} \tag{14}$$

where the constraint range $[a, b]$ corresponds to the penalty terms stipulated by relevant legal statutes.

For the task of charge prediction, the penalty function L_P is modified as follows:

$$L_P(y_{\text{pred}}, \lambda) = \begin{cases} 0 & \text{if } a \le y_{\text{pred}} \le b, \\ \lambda & \text{otherwise}, \end{cases} \tag{15}$$

where the constraint range $[a, b]$ corresponds to the charges stipulated by relevant legal statutes.

The constraint loss function, denoted as L', is formulated by summing the cross-entropy loss L with the penalty function L_P.

$$L'(y_{\text{pred}}, y_{\text{true}}, \lambda) = L(y_{\text{pred}}, y_{\text{true}}) + L_P(y_{\text{pred}}, \lambda), \tag{16}$$

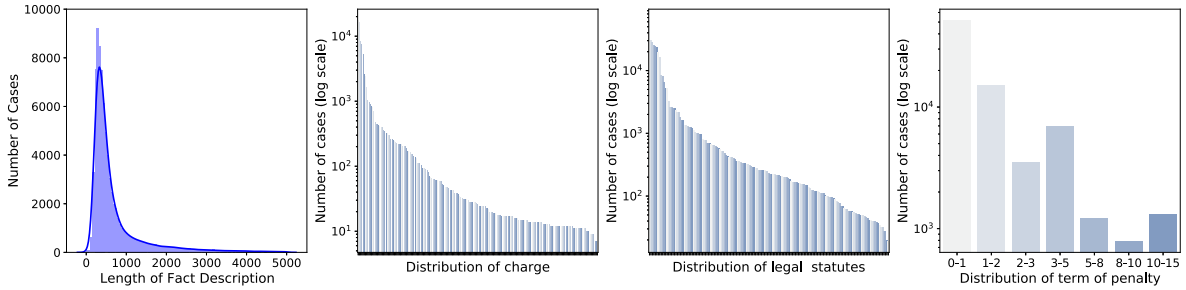

Fig. 6. Visualization of text length and label distribution in the CAIL-Long criminal dataset.

the penalty function is utilized to constrain the model's predictions. It imposes a penalty when the predicted value, y_{pred}, falls outside the constraint range $[a, b]$. The penalty is calculated by the distance of y_{pred} from the nearest limit of this range. The penalty is determined by the distance between y_{pred} and the nearest boundary of the constraint range. Therefore, the mechanism aims to predict outcomes within the constraint range, leading to more accurate and reliable outcomes.

The final loss function, L_f, integrates the losses from three sub-tasks and is represented as:

$$L_f = L_{statutes} + L'_{charge}(\lambda_1) + L'_{term}(\lambda_2). \tag{17}$$

In this formulation, $L_{statutes}$ denotes the binary cross-entropy loss for predicting legal statutes. L'_{charge} and L'_{term} represent the constrained cross-entropy losses for the prediction of the charge and the penalty term, respectively. λ_1 and λ_2 represent the penalty coefficients for the prediction of the charge and the penalty term, respectively.

4. Experiments

4.1. Datasets

Following prior work (Xiao et al., 2021), we utilized the CAIL-Long criminal dataset,[1] a publicly available collection of 1,129,053 criminal cases from China Judgments Online. Notably, the average case length of CAIL-Long criminal is much closer to real-world scenarios compared to CAIL-2018 (Xiao et al., 2021, 2018). Each case in the dataset includes a fact description, applicable legal statutes, charges, and the term of penalty. Cases involving multiple charges, which constitute less than 1% of the dataset, were removed from the study. After preprocessing, the training, validation, and test datasets comprise 80,747, 17,164, and 17,274 samples, respectively. Following Xiao et al. (2018), the term of penalty is mapped to a predefined interval, providing a range for each case's penalty term. The intervals are as follows: [0-1], [1-2], [2-3], [3-5], [5-8], [8-10], and [10-15] years. In this way, the dataset includes 199 charges, 162 legal statutes, and 7 penalty terms. Detailed statistics of the datasets, including illustrations of text length and label distribution, are shown in Fig. 6.

4.2. Evaluation metrics

Evaluation of events extraction. To assess the efficiency and applicability of event extraction, we employ both micro-averaged and macro-averaged metrics, including precision, recall, and F1 score.

Evaluation of rationale generation. Both ROUGE (Lin, 2004) and BLEU (Papineni, Roukos, Ward, & Zhu, 2002) metrics are employed to evaluate the effectiveness of rationale generation. BLEU is utilized to evaluate the similarity between the generated rationales and the original rationales, the ROUGE is used to measure the alignment between the generated rationales and the original rationales. We present the performance using ROUGE-1, ROUGE-2, and ROUGE-L,[2] as well as BLEU-1, BLEU-2, and BLEU-N.[3]

Evaluation of legal judgment prediction. Following Luo et al. (2017), accuracy (Acc.), Macro-Precision (MP), Macro-Recall (MR), and Macro-F1 (MF) are used to evaluate the performance of charge and term of penalty. We employ Precision@k (P@k) and Normalized Discounted Cumulative Gain at k (NDCG@k), where $k = \{1, 3, 5\}$ for P@k, and $k = \{5, 10, 20\}$ for NDCG@k, to evaluate the performance of legal statutes. P@k measures the proportion of relevant items among the top k predictions, while NDCG@k assesses how well the top k predictions correspond to the ideal ranking order. The formulas are as follows:

$$\text{P@}k = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{k} rel_{ij}}{k}, \tag{18}$$

$$\text{NDCG@}k = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{IDCG}_i}\sum_{j=1}^{k}\frac{2^{rel_{ij}}-1}{\log_2(j+1)}. \tag{19}$$

Here, N represents the number of test cases, rel_{ij} is the relevance score of the jth item for the ith case, and IDCG_i is the ideal discounted cumulative gain for the ith case.

4.3. Baselines

To evaluate the effectiveness of JuriSim, we select several PLMs that focus on long document processing:

- **BERT+TextRank** (Park, Vyas, & Shah, 2022): This approach truncates the first 512 tokens of a document and then selects an additional 512 tokens using TextRank,[4] an unsupervised sentence ranking algorithm. Both token sets are encoded using BERT encoders respectively. The encoded outputs are then concatenated and processed through a fully-connected layer for document classification.
- **BERT+Random** (Park et al., 2022): This approach truncates the first 512 tokens of a document, and then selects a random selection of up to 512 tokens from the remaining text. Other Setting remains consistent with the BERT+TextRank.
- **ToBERT** (Pappagari et al., 2019; Park et al., 2022): A hierarchical approach is adopted to handle long documents. The documents are segmented into multiple paragraphs, each with a maximum length of 128 tokens. These segmented paragraphs are then processed using a Transformer layer over BERT-based sentence representations.
- **Longformer** (Beltagy et al., 2020): Longformer is utilized for processing long texts, employing an attention mechanism with linear complexity.

[1] The dataset can be downloaded from https://data.thunlp.org/legal/CAIL-Long.tar.gz.

[2] https://github.com/pltrdy/rouge.

[3] BLEU-N is the average of BLEU-1, BLEU2, BLEU-3 and BLEU-4.

[4] https://github.com/letiantian/TextRank4ZH.

- **Lawformer** (Xiao et al., 2021): Lawformer is used for processing long Chinese legal documents. It was pre-trained on a large corpus of extensive Chinese legal cases, employing a Longformer-based architecture (Beltagy et al., 2020).

To further compare the effectiveness of JuriSim and LLM-based methods, we also selected the following representative LLMs.

- **Qwen2** (Yang et al., 2024): Qwen2 series of LLMs, pre-trained on high-quality multilingual datasets, exhibit strong multilingual comprehension. We select Qwen2-7B-Instruct[5] and Qwen2-72B-Instruct[6] as our baseline.
- **GPT-3.5-turbo** (Brown et al., 2020): GPT-3.5-turbo, an iteration of the GPT series enhanced for speed and efficiency, serves as a baseline and can understand and generate complex natural language.
- **GPT-4-turbo** (Achiam et al., 2023): GPT-4-turbo, an iteration of the GPT series, leverages deeper layers and a larger dataset, enhancing its capacity for complex legal reasoning and multilingual capabilities.

To compare the effectiveness of JuriSim and state-of-the-art LJP methods, we selected the following representative baselines:

- **NeurJudge_Lawformer** : NeurJudge (Yue, Liu, Jin et al., 2021) separates the fact description into distinct circumstances, utilizing these to predict legal judgments such as charges, law statutes, and term of penalties. Diverging from the original model, we replaced the Bi-GRU encoder with Lawformer, which is more adapted for processing lengthy legal texts.
- **RLJP_Lawformer** : RLJP (Wu et al., 2022) processes the LJP task in two steps: firstly generating rationales from fact descriptions, followed by predicting the judgment utilizing both the facts and the generated rationales. Unlike the original model, we replaced the Transformer encoder with Lawformer, which can be used to process lengthy legal texts.
- **ADAPT** (Deng et al., 2024): ADAPT employs a multi-step reasoning prompt learning method for LJP: it first focuses on decomposing case facts, discriminating among possible charges, and synthesizing a final judgment. The method fine-tunes the Qwen-7B-Instruct model using a multi-task learning strategy.

4.4. Implementation details

In our experiment setup, we utilized the Hugging Face Transformers[7] and PyTorch Lightning[8] for model development and training. For all baselines and JuriSim models, due to the limitations of GPU memory, we employed FP16 mixed precision fine-tuning to optimize memory usage during training. The AdamW optimizer (Loshchilov & Hutter, 2017) was used to fine-tune each model on the CAIL-Long criminal dataset, with the learning rate set to 2e−5 and a linear rate scheduler for learning adjustments. The models were trained for 10 epochs on an NVIDIA A100 40G GPU, with a batch size of 8.

For the JuriSim framework, Chinese-RoBERTa-wwm-ext[9] is used to initialize the encoder for both the events and the rationales. The maximum input lengths for the events encoder and the rationales encoder are set to 510 and 200, respectively. JuriSim_ToBERT, JuriSim_Longformer, and JuriSim_Lawformer use ToBERT, Longformer, and Lawformer, respectively, to encode the fact description. The maximum input lengths and PLMs weights for initializing the fact description encoder are presented in Table 1. The number of heads m in the dual residual cross-attention mechanism is set to 12. Both β_1 and β_2 in the dual residual cross-attention mechanism are set to 0.1, and a detailed sensitivity analysis of these hyperparameters is discussed in Section 5.4. The fully connected layer (FC) consists of two linear layers: the size of the first linear layer is set to 1024, while the size of the second linear layer depends on the size of the task label, set respectively at 162, 199, and 7 for predicting legal statutes, charges, and penalty terms. For the constraint loss function L'_{charge} and L'_{term}, λ_1 and λ_2 are set to 0.5. Each legal statute is associated with a corresponding subset of charges and terms, as described in Section 1, where we can manually extract the constraint range $[a, b]$ for charges and terms of penalties from the descriptions of legal statutes, respectively.

To implement BERT+TextRank, BERT+Random, ToBERT, Longformer, Lawformer, NeurJudge_Lawformer, and RLJP_Lawformer, the maximum input length and PLM weights for initializing the fact description encoder are presented in Table 1. BERT+TextRank and BERT +Random first select the initial 512 tokens from the fact description. For parts exceeding 512 tokens, TextRank and random strategies are used to select an additional 512 tokens, respectively. These tokens are then fed into a shared BERT model, and the representations are used to predict the legal judgment. Longformer and Lawformer take the fact description as input and use the pooled output as the fact representation to predict the legal judgment. For NeurJudge_Lawformer and RLJP_Lawformer, we replace the BI-GRU encoder with Lawformer for encoding the fact description, and reproduce their code.[10,11]

We implement Qwen2, GPT-3.5-turbo, and GPT-4-turbo using zero-shot learning with the instruction prompt "*Please analyze the following case to determine the possible charges against the defendant, the applicable legal statutes, and the term of penalty.*" ("请对以下案件进行分析, 确定被告人可能面临的罪名, 适用的法律条文, 以及预计的刑期。") to generate the legal judgment. Due to limited budget, we randomly selected 3000 samples from the CAIL-Long criminal test dataset to evaluate the performance of GPT-3.5-turbo and GPT-4-turbo. To implement ADAPT (Deng et al., 2024), we use the fine-tuned ADAPT model[12] with the instruction prompt *"Please use the ADAPT framework to analyze the possible charges against the defendant, the applicable legal provisions, and the term of penalty"*. ("请你采用 ADAPT 框架分析以上案件中该被告人可能被判处的罪名 适用法条和刑期"). Consistent with Deng et al. (2024), greedy decoding is used for all generative models. For any generated charges not present in the label pool, BGE (Xiao, Liu, Zhang, & Muennighof, 2023) is employed to map them to the closest charge in the pool based on their representations.

4.5. Events extraction implementation

Experiments conducted on a publicly available LEVEN dataset (Yao et al., 2022) to evaluate the efficiency and applicability of our events extraction model. The BERT-CRF model was initialized using bert-base-chinese.[13] The learning rate for the model was set to 3e−5, and the CRF learning rate was set to 1e−3. It was trained with a batch size of 24 over 4 epochs. During training, the input length was set to 128 tokens, which was increased to 512 tokens for the inference phase. Similarly, the DMBERT model was also initialized with bert-base-chinese, with the input length set to 512 tokens. The learning rate was set to 5e−5, the batch size to 32, and it ran for 4 epochs.

The event extraction performance of the model is demonstrated as follows: on a micro level, the precision, recall, and F1 score are 83.64%, 86.37%, and 84.99%, respectively. On a macro level, the precision, recall, and F1 score are 81.41%, 81.30%, and 80.10%, respectively.

[5] https://huggingface.co/Qwen/Qwen2-7B-Instruct.
[6] https://huggingface.co/Qwen/Qwen2-72B-Instruct.
[7] https://huggingface.co/transformers/.
[8] https://lightning.ai/docs/pytorch/1.6.0/.
[9] https://huggingface.co/hfl/chinese-roberta-wwm-ext.
[10] https://github.com/yuelinan/NeurJudge.
[11] https://github.com/wuyiquan/RLJP.
[12] https://huggingface.co/ChenlongDeng/ADAPT-Qwen2-7B-CAIL2018-step-8765.
[13] https://huggingface.co/bert-base-chinese.

Table 1
Parameters for the model implementation.

Models	PLMs	Input length
BERT+TextRank	https://huggingface.co/bert-base-chinese	1024
BERT+Random	https://huggingface.co/bert-base-chinese	1024
ToBERT	https://huggingface.co/bert-base-chinese	2560
Longformer	https://huggingface.co/ValkyriaLenneth/longformer_zh	2000
Lawformer	https://huggingface.co/thunlp/Lawformer	2000
NeurJudge_Lawformer	https://huggingface.co/thunlp/Lawformer	2000
RLJP_Lawformer	https://huggingface.co/thunlp/Lawformer	2000
JuriSim_ToBERT	https://huggingface.co/bert-base-chinese	2560
JuriSim_Longformer	https://huggingface.co/ValkyriaLenneth/longformer_zh	2000
JuriSim_Lawformer	https://huggingface.co/thunlp/Lawformer	2000

Table 2
Performance comparison of rationale generation models.

Models	BLEU-1	BLEU-2	BLEU-N	ROUGE-1	ROUGE-2	ROUGE-L
C3VG	52.10	43.50	40.60	60.10	40.50	62.50
MT5-Large	64.39	59.29	57.72	73.13	62.01	72.46

4.6. Rationale generation implementation

Experiments conducted on the dataset from Yue, Liu, Wu et al. (2021) to evaluate the efficiency and applicability of rationale generation. The rationale generation model was initialized with the pre-trained MT-large[14] (Xue et al., 2021). For training, we set the input and output lengths of the models to 700 and 200 tokens, respectively. During inference, the input and output lengths were set to 1300 and 200 tokens. We fine-tuned the model using the AdamW optimizer (Loshchilov & Hutter, 2017) with a constant learning rate of 10^{-4}, over 5 epochs and a batch size of 16. In our decoder strategy, all settings were kept consistent with He et al. (2023).

As shown Table 2, we compared MT5-Large model with the C3VG method (Yue, Liu, Wu et al., 2021). Inspired by PLMs, MT5-Large model outperformed the C3VG method by achieving a 42.17% BLEU-N and a 15.94% ROUGE-L score, demonstrating the efficacy of MT5-Large model.

5. Experiment results and analysis

5.1. Comparison with state-of-the-art

We conducted a comparison of JuriSim-based methods with two BERT-based models: BERT+Random and BERT+TextRank. Additionally, we evaluated three PLMs that focus on long document processing. Furthermore, our method was compared against two state-of-the-art LJP models, NeurJudge and RLJP. For a fair comparison, both NeurJudge and RLJP utilize Lawformer as the fact encoder. We also compared four LLMs, as well as ADAPT, which fine-tunes the Qwen-7B-Instruct model with legal corpora.

The experiment results are presented in Tables 3 and 4. We observed that JuriSim_Lawformer consistently outperforms other baseline methods and state-of-the-art LJP models on most metrics. This demonstrates the effectiveness of our proposed method in predicting legal judgments. We then analyzed the strengths and weaknesses of various methods from different perspectives.

Long document models vs. BERT-based models: The results indicate that BERT+TextRank consistently outperforms BERT+Random on most metrics, especially in predicting charge and term of penalty. This suggests that the extraction of key text from fact description enhance the performance of LJP. Moreover, our findings indicate that the truncation process omits crucial information, resulting in a decline in classification performance.

Models based on Resource-Efficient Transformers, such as Lawformer and Longformer, demonstrate superior performance over BERT+TextRank on most metrics. This illustrates the significant advantage of Resource-Efficient Transformers-based models in the task of LJP for long Chinese legal documents. In contrast, ToBERT, a model based on Hierarchical Transformers, considerably underperforms compared to BERT+TextRank. This suggests that Hierarchical Transformers-based models is not be suitable for this task. One possible reason could be that segmenting the fact description into multiple parts might disrupt the structural completeness of legal texts, impacting model performance.

Furthermore, BERT+TextRank achieves competitive performance on the CAIL-Long criminal dataset. We guess that this could be attributed to the fact that over 50% of the fact description in the CAIL-Long criminal dataset are shorter than 10,000 tokens, potentially enhancing the effectiveness of BERT+TextRank. To validate this hypothesis, a detailed analysis and discussion of the effect of fact description length on model performance is presented in Section 5.3.

Longformer/Lawformer vs. Hierarchical Transformers: To evaluate the efficacy of Resource-Efficient Transformers (e.g., Lawformer, Longformer) and Hierarchical Transformers (e.g., ToBERT) in representing long documents, we conducted experiments in LJP tasks using various transformers. The results indicate that Longformer generally outperforms ToBERT, suggesting that Resource-Efficient Transformers are more effective in processing long documents, especially within the Chinese legal domain. This superiority may be due to the structural completeness of legal texts, which is distinct from general texts. The common method of segmenting legal texts could disrupt their structural integrity, consequently diminishing the representational capacity of fact descriptions. Furthermore, Lawformer outperforms Longformer, indicating the effectiveness of continued pre-training in the legal domain.

JuriSim-based methods vs. Long document models: As illustrated in Tables 3 and 4, the performance of JuriSim-based methods consistently surpass other long document models, including ToBERT, Longformer, and Lawformer. In particular, JuriSim_Lawformer shows a relative improvement of 3.45% and 3.05% in MF for predicting charges and term of penalties, respectively, compared to Lawformer.

JuriSim-based methods vs. State-of-the-art LJP models: JuriSim_Lawformer model significantly outperforms state-of-the-art LJP models. In particular, compared to RLJP_Lawformer, JuriSim_Lawformer demonstrates relative improvements of 4.57%, 16.05%, and 5.79% in NDCG@5 for legal statutes prediction, MF for charge prediction, and MF for term of penalty prediction, respectively. Furthermore, compared against NeurJudge_Lawformer, JuriSim_Lawformer shows relative improvements of 0.87%, 7.11%, and 5.08% for the same metrics. These improvements can be attributed to the JuriSim framework's ability to combine the knowledge of judicial trial logic. The JuriSim framework effectively integrates events, rationales, and the topological relationships between charges, terms, and applicable law statutes. This approach significantly improves the model's ability to capture key information, such as criminal behaviors and consequences. In addition, JuriSim introduces a residual cross-attention mechanism, enabling the model to focus on key information and reduce the loss of important

[14] https://huggingface.co/google/mt5-large.

Table 3
Performance comparison of various models on legal statutes across P@K and NDCG@K metrics. Bolded numbers represent the best performance. P@1 is reported for LLMs since these models only generate a legal statute.

Models	Legal statutes					
	P@1	P@3	P@5	NDCG@5	NDCG@10	NDCG@20
BERT-based models						
BERT+Random	95.51	76.77	62.97	86.17	90.24	92.16
BERT+TextRank	95.60	76.70	62.83	86.06	90.16	92.08
Long document models						
ToBERT	95.31	76.26	62.45	85.60	89.68	91.71
Longformer	95.72	76.73	62.89	86.12	90.17	92.08
Lawformer	95.95	**77.24**	63.22	86.53	90.38	92.25
State-of-the-art LJP models						
NeurJudge_Lawformer	95.72	76.55	62.67	85.89	89.95	91.93
RLJP_Lawformer	92.11	74.53	60.76	82.85	87.18	89.42
Large language models						
Qwen2-7B	88.25	–	–	–	–	–
Qwen2-72B	91.84	–	–	–	–	–
GPT-3.5-turbo	12.36	–	–	–	–	–
GPT-4-turbo	48.68	–	–	–	–	–
ADAPT	81.85	–	–	–	–	–
Ours						
JuriSim_ToBERT	95.86	77.16	63.14	86.41	90.28	92.16
JuriSim_Longformer	95.82	76.98	62.89	86.16	90.24	92.13
JuriSim_Lawformer	**96.10**	77.17	**63.28**	**86.64**	**90.59**	**92.45**

Table 4
Comparison of JuriSim with BERT-based models, long document models and state-of-the-art LJP models on charges and term of penalties.

Models	Charge				Term of penalty			
	Acc.	MP	MR	MF	Acc.	MP	MR	MF
BERT-based models								
BERT+Random	95.09	74.36	69.15	69.85	75.19	53.63	50.77	51.80
BERT+TextRank	95.36	77.84	71.66	72.81	74.86	54.04	51.10	52.35
Long document models								
ToBERT	94.88	76.54	70.75	71.70	74.39	53.93	50.51	51.60
Longformer	95.50	75.84	72.07	71.72	75.66	54.07	52.26	52.84
Lawformer	95.67	79.60	74.22	74.87	75.80	55.06	51.92	53.38
State-of-the-art LJP models								
NeurJudge_Lawformer	95.34	76.99	71.35	72.31	74.99	53.68	51.40	52.35
RLJP_Lawformer	95.10	74.58	64.39	66.74	75.07	53.76	50.54	52.00
Large language models								
Qwen2-7B	88.47	64.12	53.60	53.52	33.94	23.08	23.36	20.26
Qwen2-72B	92.28	75.74	70.27	68.55	40.99	28.30	34.05	26.85
GPT-3.5-turbo	82.40	38.04	39.58	35.40	32.86	25.04	15.01	15.73
GPT-4-turbo	86.63	43.58	45.64	41.90	36.15	23.04	17.45	19.33
ADAPT	86.30	63.63	**77.63**	65.35	55.29	38.39	26.57	30.54
Ours								
JuriSim_ToBERT	95.54	79.01	73.80	74.54	75.46	55.00	50.63	52.57
JuriSim_Longformer	95.80	77.82	76.64	75.73	75.28	54.53	52.87	53.59
JuriSim_Lawformer	**96.05**	**81.28**	77.26	**77.45**	**76.10**	**56.09**	**54.16**	**55.01**

information as well as the retention of incorrect information during the aggregation process.

Large language models: Qwen2-7B, Qwen2-72B, and ADAPT significantly outperform GPT-3.5-turbo and GPT-4-turbo across various metrics. This is because Qwen2-7B, Qwen2-72B, and ADAPT incorporate more Chinese corpora, achieving competitive performance in the Chinese domain. Additionally, ADAPT outperforms Qwen2-7B and Qwen2-72B in most metrics, especially in the task of term of penalty, mainly because ADAPT employs a multi-step reasoning prompt learning method to improve LJP performance. Furthermore, we found that LLMs perform worse than PLMs in most metrics, primarily because LLMs do not excel in prediction tasks. However, LLMs offer stronger interpretability compared to PLMs, which is more applicable in real-world scenarios.

JuriSim-based methods vs. Large language models: Compared to LLMs, JuriSim_Lawformer achieves a more significant advantage in all metrics except MR in charge prediction. Our improvements suggest that even though LLMs have strong generative capabilities, they do not perform well in prediction tasks. We found that ADAPT and JuriSim-based methods achieved competitive performance in the charge prediction task, primarily because charge labels have actual meaning, while performance significantly declines when the label has no actual meaning (e.g., the index of the law article and prison term). Therefore, using LLMs for LJP remains a more challenging direction.

5.2. Ablation studies

We conducted experiments on the ablated versions of JuriSim_Lawformer to analyze the impact of different modules to the performance of JuriSim. The results are presented in Table 5.

Constraint Loss. To verify the efficacy of constraint loss, we replaced the constraint cross-entropy loss with the standard cross-entropy

Table 5
Performance comparison of various JuriSim_Lawformer model variants on legal statutes, charges and term of penalties. DRCA is the abbreviation for Dual Residual Cross-Attention Mechanism.

Models	Legal statutes			Charge				Term of penalty			
	P@1	P@3	NDCG@5	Acc.	MP	MR	MF	Acc.	MP	MR	MF
JuriSim_Lawformer	96.10	77.17	86.64	96.05	81.28	77.26	77.45	76.10	56.09	54.16	55.01
w/o constraint loss	96.09	77.26	86.67	96.05	80.83	77.29	77.30	75.93	57.02	54.89	55.91
w/o rationales	95.75	77.33	86.65	95.93	80.96	77.23	77.08	75.54	55.11	52.58	53.75
w/o DRCA	95.87	77.66	86.95	95.54	78.50	71.84	73.56	75.65	54.97	51.89	53.30
w/o residual mechanism	95.43	75.42	85.04	96.08	81.12	76.92	77.42	75.97	55.43	55.79	55.48
Lawformer+events	95.92	77.71	86.89	95.90	79.85	76.93	76.77	75.99	55.78	53.31	54.48
Lawformer	95.95	77.24	86.53	95.67	79.60	74.22	74.87	75.80	55.06	51.92	53.38

loss. A comparative analysis between JuriSim_Lawformer and JuriSim_Lawformer (w/o constraint loss) indicated a decrease in the charge prediction MF by 0.15%, while the term of penalty prediction MF increased by 0.90%.

The removal of the constraint loss function resulted in diminished performance in charge prediction but led to an improvement in term of penalty prediction. These results suggest that the constraint loss function is crucial for optimizing charge prediction tasks, with a potential adverse impact on term of penalty prediction tasks. This could be attributed to over 60% of legal statutes providing varied penalty ranges based on the severity of the crime, making it challenging to predict term of penalty solely based on the applicable legal statutes. This finding indicates a potential future research direction, which might involve incorporating the severity of the defendant's criminal behavior to enhance the accuracy of penalty term predictions.

Rationales. Then, we investigated the effect of rationales on JuriSim_Lawformer's performance. A comparison between JuriSim_Lawformer (w/o rationales) and JuriSim_Lawformer indicates a slight decrease in legal statutes prediction, a decrease in charge prediction MF by 0.37%, and a more significant decrease of 1.26% in term of penalty prediction MF. These results suggest that rationales enhances the model's capability to capture the key behaviors and consequences, thereby improving the performance of LJP.

Events. We conducted a comparative analysis between Lawformer +events and the Lawformer to evaluate the impact of introducing events to the model. Our results show that integrating events leads to improvements in legal statutes prediction, as indicated by a 0.47% increase in P@3 and a 0.36% increase in NDCG@5. Moreover, we observed significant improvements in charge prediction MF and term of penalty prediction MF, with increases of 1.60% and 1.10% respectively. These findings indicate that incorporating events into the model can improve the location of key events in fact descriptions, thereby enhancing the overall performance of LJP.

Dual Residual Cross-Attention Mechanism. An ablation study was conducted to verify the effectiveness of the Dual Residual Cross-Attention Mechanism. A significant decrease was observed in both the charge prediction MF and the term of penalty prediction MF when the dual residual cross-attention mechanism was removed. This highlights the importance of the dual residual cross-attention mechanism in effectively integrating and processing the different components of legal documents.

Residual Mechanism. An ablation study was conducted to verify the effectiveness of the Residual Mechanism. We found that removing the Residual Mechanism led to a significant decline in performance metrics for law article and charge tasks, with a slight increase in the MF for the term of penalty. This indicates that the Residual Mechanism allows the model to reduce the loss of important information and the retention of incorrect information during the aggregation process, thereby enhancing the performance of LJP.

5.3. Long text performance

We further investigated the JuriSim_Lawformer model's performance on lengthy texts. The length of the fact descriptions is divided into three groups: the first group with texts exceeding 1000 tokens (labeled as "length > 1000"), the second group with lengths between 500 and 1000 (labeled as "1000 ≥ length > 500"), and the third group with lengths less than 500 tokens, (labeled as "length ≤ 500"). The performance of BERT+TextRank, ToBERT, Lawformer, and JuriSim_Lawformer was compared on these groups.

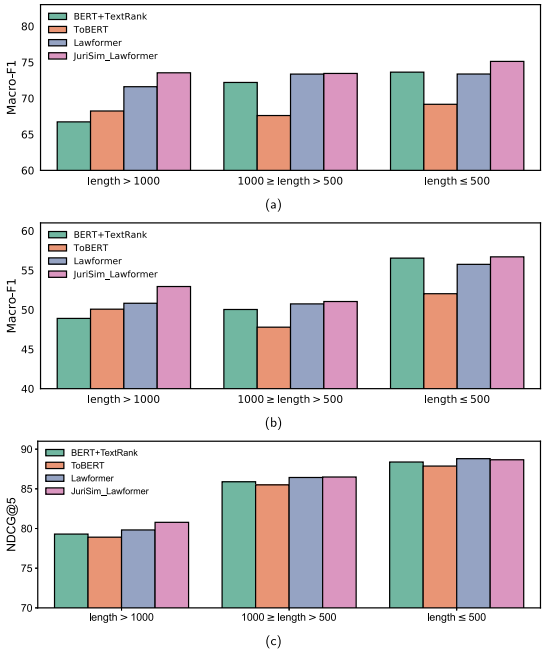

Fig. 7. (a) Comparison of macro-F1 scores for charge prediction across different fact description length groups. (b) Comparison of macro-F1 scores for term of penalty prediction across different fact description length groups. (c) Comparison of NDCG@5 scores for legal statutes prediction across different fact description length groups.

Fig. 7 illustrate the performance of different models when fact descriptions of different lengths are used in charge prediction, term of penalty prediction, and legal statutes prediction, respectively.

Notably, in the "length > 1000" group, the JuriSim_Lawformer model significantly outperforms other models. Specifically, JuriSim_Lawformer surpasses Lawformer by 1.93%, 2.12%, and 1.11% in MF for charge prediction, MF for term of penalty prediction, and NDCG@5 for legal statutes prediction, respectively. Furthermore, in the "length ≤ 500" group, the BERT+TextRank model shows competitive results, highlighting its effectiveness in shorter legal cases. By contrast, Lawformer performs worse than the BERT+TextRank in shorter legal cases. However, the JuriSim_Lawformer model still outperforms the BERT+TextRank. These findings not only validate the robustness of our method in LJP tasks but also further demonstrate the capability of the JuriSim method in handling long legal documents.

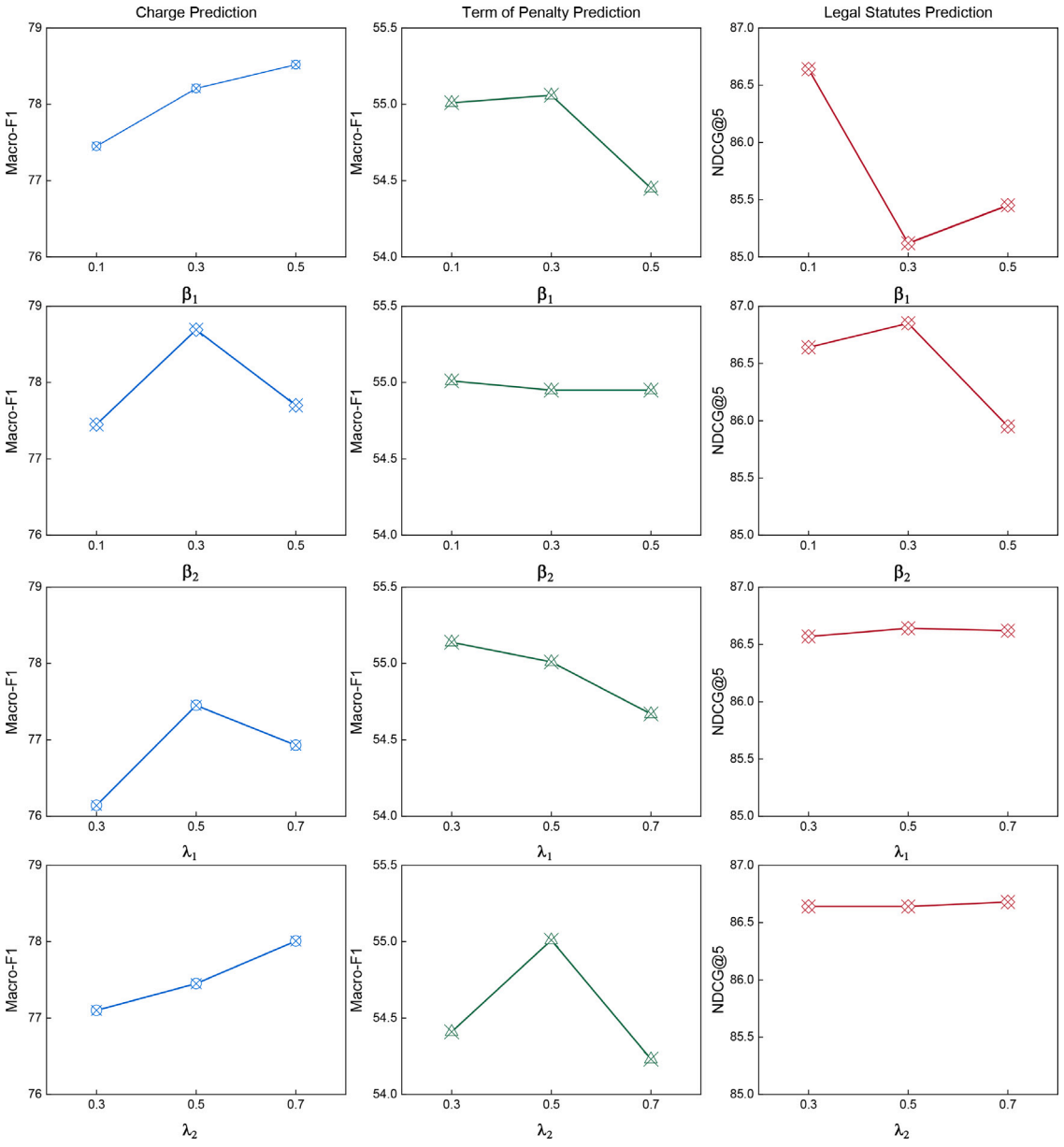

Fig. 8. Sensitivity analysis of hyperparameters β_1, β_2, λ_1, and λ_2 in the JuriSim_Lawformer model, demonstrating the performance of charge prediction, term of penalty prediction, and legal statutes prediction.

5.4. Sensitivity analysis of hyperparameters

In this section, we conducted a sensitivity analysis of the hyperparameters in our JuriSim_Lawformer model, focusing particularly on the dual residual cross-attention mechanism and the constrained cross-entropy loss. Our analysis investigated how the hyperparameters β_1, β_2, λ_1, and λ_2 affect the model's performance. Fig. 8 illustrates the impact of these hyperparameters on the JuriSim_Lawformer model's performance across different LJP tasks.

Impact of β_1 and β_2: The hyperparameters β_1 and β_2 in JuriSim_Lawformer influence the event and rationale representations within the dual residual cross-attention mechanism. We observed that increasing β_1 from 0.1 to 0.3 led to significant improvements in charge prediction, but it also resulted in a noticeable decrease in legal statutes prediction. Reducing β_1 enhances the stability across tasks. Similarly, increasing β_2 from 0.1 to 0.3 caused an improvement in charge prediction tasks, while a slight decrease on term of penalty and legal statutes predictions. However, further increasing β_2 from 0.3 to 0.5 led to a decline in performance on all tasks. This suggests that β_2 is highly sensitive to parameter adjustments, and larger adjustments affect the performance of term of penalty prediction and legal statutes prediction tasks.

Influence of λ_1 and λ_2: Furthermore, we evaluate the impact of λ_1 and λ_2, the hyperparameters in the constrained cross-entropy loss. Increasing λ_1 from 0.3 to 0.5 led to a significant improvement in charge prediction, with increases of 1.72% in MF, but resulted in a slight decrease in term of penalty prediction. However, further increasing λ_1 to 0.7 led to a decrease in performance for predicting charges and term of penalties. Similarly, increasing λ_2 from 0.3 to 0.5 demonstrated an improvement in term of penalty, with increases of 1.10% in MF. However, as λ_2 was further increased, the performance in term of penalty significantly declined, despite a slight improvement in charge prediction. Adjusting λ_1 and λ_2 within an effective range enhanced the performance of both charge and term of penalty predictions, demonstrating the effectiveness of our constrained loss function.

Fact Description	... the defendants Li Mou1, Li Mou2 (already sentenced), and Li Mou3 (already sentenced) assaulted Zheng Mou, Feng Mou1, and Feng Mou2 in their residence with sticks and kitchen knives because they had unsuccessfully demanded wages from the victim Zheng Mou. After the attack, they fled in a taxi. Zheng Mou suffered a complete fracture of the right scaphoid bone, classified as a minor injury level two, and had a total scar length of 34.5cm, also classified as a minor injury level two, plus a scalp hematoma classified as a slight injury. Feng Mou1 suffered a left frontal epidural hematoma, classified as a minor injury level one, a skull fracture classified as a minor injury level two, and bruises over an area of 202.0 square centimeters, classified as a slight injury. Feng Mou2 suffered a brain contusion and laceration classified as a minor injury level one, a fracture of the left ulna and scars totaling 15.5cm, both classified as minor injuries level two, and a 6.5cm neck laceration and fractures of the second proximal phalange of the left hand and the third and fourth metacarpal bones, all classified as slight injuries. ...			
Events	**Events**		**Event Types**	
	demanded → in their residence → assaulted → fled → fracture → fracture → contusion → laceration → ...		requesting/demanding → entering a dwelling → personal injury → escape → injury → injury → injury → injury → ...	
Rationale	The court determined that Li Mou1 disregarded national laws, intentionally harmed others, causing minor injuries to multiple people, and thus committed a crime. This case involved a joint offense, with Li Mou1 playing a major role and being the principal offender. Accordingly, he should be punished for all the crimes he participated in. After committing the crime, Li Mou1 voluntarily surrendered to the police and truthfully confessed, qualifying for a lighter punishment according to law. Li Mou1's family reached a settlement with the victims Zheng Mou and Feng Mou1, compensating Feng Mou2 5,000 yuan. Feng Mou2 issued a letter of forgiveness for Li Mou1, which allows for a discretionary lighter punishment according to law.			
Judgment		**Ground Truth**	**Lawformer**	**JuriSim_Lawformer**
	Charge	Crime of intentional injury	Crime of picking quarrels and provoking trouble ❌	Crime of intentional injury ✅
	Term of Penalty	1~2 years	1~2 years ✅	1~2years ✅

Fig. 9. Case study illustrating the prediction of crime of intentional injury.

Fact Description	... Yuan and Hang (handled in a separate case) contacted the defendant Cao, requesting to buy heroin. Cao carried the drugs from Liangping County, Chongqing City to room 418 of the "Smaer Hotel" on Shixian Road, Wanzhou District, where he handed over a package of heroin to Hang. After Hang left, Yuan handed over 7,000 RMB of drug money to Cao. Following the completion of the transaction, the police apprehended Cao and seized the drug money amounting to 7,000 RMB and 21.07 grams of drugs. The Wanzhou District Public Security Bureau's Forensic Evidence Identification Department confirmed that the seized substance contained heroin. ...			
Events	**Events**		**Event Types**	
	contacted → requesting → buy → handed over → apprehended → seized → ...		contact/communication → request/ask → buy → pay/give → arrest → search/seize → ...	
Rationale	The court believes that Cao knowingly sold heroin, a narcotic, to others, and his actions constitute a crime. The defendant Cao truthfully confessed to his crimes and can be given a lighter punishment according to the law.			
Judgment		**Ground Truth**	**Lawformer**	**JuriSim_Lawformer**
	Charge	Crimes of smuggling, trafficking, transporting and manufacturing drugs	Crime of illegal drug possession ❌	Crimes of smuggling, trafficking, transporting and manufacturing drugs ✅
	Term of Penalty	7~10 years	2~3 years ❌	7~10years ✅

Fig. 10. Case study on the prediction of crimes of smuggling, trafficking, transporting, and manufacturing drugs.

5.5. Case study

This section presented two illustrative cases to demonstrate the efficiency of the JuriSim framework in handling LJP for complex legal cases.

A legal case of the crime of intentional injury is illustrated in Fig. 9. The fact description details that *"the defendant, Li, along with his accomplices, entered Zheng and Feng's residence armed with clubs and knives. They assaulted them over a wage dispute, causing injuries of varying severity, including fractures and contusions"*. Lawformer incorrectly predicted the charge as "crime of picking quarrels and provoking trouble", potentially misconstruing the actions as merely provocative and harassing behavior. In contrast, JuriSim_Lawformer, by analyzing the events and event types, such as "demanded", "entering a dwelling", "assaulted", and "injury", along with rationales like *"Li's disregard for the law and intentional harm to others"*, identifies the legal case as an intentional injury related to an unresolved wage dispute. Consequently, JuriSim_Lawformer correctly predicts the charge as "crime of intentional injury".

Fig. 10 presents a legal case involving crimes of smuggling, trafficking, transporting, and manufacturing drugs. The facts describe *the defendant, Cao, suspected of selling heroin to Hang and Yuan on September 12, 2014. Following Cao's arrest, the police found 21.07 grams of heroin and 7000 yuan.* Lawformer inaccurately predicted the charge as "crime of illegal drug possession", possibly due to its focus on the drug-related aspects without considering the actions of trafficking. However, JuriSim, utilizing its event analysis capabilities, identifies key events in the case, such as "buy" and "handed over". The rationale, which involves *"Cao knowingly selling heroin to others"*, leads to a prediction of "crimes of smuggling, trafficking, transporting, and manufacturing drugs". Additionally, as Lawformer interpreted the case as drug possession rather than trafficking, it also inaccurately predicted the term of penalty.

These case studies highlight the significance of comprehending events and rationales in legal cases. The capacity of JuriSim to analyze these key criminal behaviors and events demonstrates its effectiveness in LJP tasks.

6. Conclusion and future work

In this paper, we present JuriSim, a novel framework for predicting Chinese criminal legal judgments by integrating the knowledge of judicial trial logic. This framework incorporates events to predict applicable legal statutes and then integrates events and rationales to predict charges and penalty terms. This approach enhances the model's ability to capture decisive information that influences judgment predictions, such as criminal events, behaviors, and consequences. Furthermore, we propose a dual residual cross-attention mechanism to interactively process facts, events, and rationales. This mechanism allows the model to reduce the loss of important information and the retention of incorrect information during the aggregation process. In addition, we introduce a constrained cross-entropy loss, leveraging the topological relationship between applicable law statutes, charges, and terms of penalties. Our experiments on the CAIL-Long criminal dataset illustrate that JuriSim outperforms state-of-the-art methods in predicting legal judgments, especially in cases with long documents.

Our future work will focus on two main areas: (1) Incorporating the severity of the criminal behavior to enhance the accuracy of penalty term prediction. (2) Investigating methods that improve the performance of similar case retrieval, further refining the framework of simulating judicial trial logic, improving the applicability of the legal system to real-world scenarios.

7. Ethical discussion

With the increasing adoption of artificial intelligence technology in the judicial field, the public is becoming more concerned about the ethical issues these systems may raise. Particularly, any small error or bias in the system could lead to significant consequences. Therefore, while our research explores models to predict legal judgments, we clearly recognize that these technologies are currently not suitable for independent use without human supervision. Our system is designed as an auxiliary tool to provide judges with data-driven insights, helping them make more informed decisions rather than replacing the judge's authority. During the implementation of this technology, ensuring the fairness and transparency of the law remains the responsibility of human judges. In the future, we plan to conduct further research to identify and eliminate potential biases in the algorithms, ensuring the practicality and fairness of these technologies.

8. Limitations

In this section, we discuss the limitations of our works as follow:

- The current benchmark datasets for Legal Judgment Prediction (LJP) models lack crucial elements such as contradictory argument data, jurisprudence, and precedents. These omissions restrict our ability to fully simulate complex real-world judgment logic and integrate real-world legal scenarios. Expanding these datasets would significantly enhance the realism and applicability of LJP models, allowing for a more comprehensive exploration of these critical areas.
- Our method is currently only applicable to Chinese criminal cases. It is worthwhile to explore more generalizable methods that can be applied to different legal systems.

CRediT authorship contribution statement

Congqing He: Conceptualization, Methodology, Software, Writing – original draft. **Tien-Ping Tan:** Supervision, Writing – review & editing. **Sheng Xue:** Writing – review & editing. **Yanyu Tan:** Writing – review & editing, Funding acquisition.

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

The dataset can be downloaded from https://data.thunlp.org/legal/CAIL-Long.tar.gz.

Acknowledgments

This work was supported by the Key Scientific Research Project of Hunan Provincial Department of Education, China (20A123), and the Key Special Education Project of Hunan Provincial Social Science Foundation, China (18ZDJ01).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F., & Sanghai, S. (2023). GQA: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245.

Alali, M., Syed, S., Alsayed, M., Patel, S., & Bodala, H. (2021). JUSTICE: A benchmark dataset for supreme court's judgment prediction. arXiv preprint arXiv:2112.03414.

Almuzaini, H. A., & Azmi, A. M. (2023). Tasbeeb: A judicial decision support system based on deep learning framework. *Journal of King Saud University-Computer and Information Sciences*, *35*(8), Article 101695.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. arXiv:2004.05150.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances In Neural Information Processing Systems*, *33*, 1877–1901.

Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural legal judgment prediction in english. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4317–4323). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1424, URL: https://aclanthology.org/P19-1424.

Chen, H., Cai, D., Dai, W., Dai, Z., & Ding, Y. (2019). Charge-based prison term prediction with deep gating network. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 6362–6367). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1667, URL: https://aclanthology.org/D19-1667.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:2004.13922.

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3504–3514.

Dai, X., Chalkidis, I., Darkner, S., & Elliott, D. (2022). Revisiting transformer-based models for long document classification. arXiv preprint arXiv:2204.06683.

Deng, C., Mao, K., Zhang, Y., & Dou, Z. (2024). Enabling discriminative reasoning in large language models for legal judgment prediction. arXiv preprint arXiv:2407.01964.

Feng, Y., Li, C., & Ng, V. (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)* (pp. 648–664). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.48, URL: https://aclanthology.org/2022.acl-long.48.

He, C., Tan, T.-P., Xue, S., & Tan, Y. (2023). Explaining legal judgments: A multitask learning framework for enhancing factual consistency in rationale generation. *Journal of King Saud University-Computer and Information Sciences*, Article 101868.

Hu, Z., Li, X., Tu, C., Liu, Z., & Sun, M. (2018). Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th international conference on computational linguistics* (pp. 487–498).

Huang, J. H., & Powers, D. (2003). Chinese word segmentation based on contextual entropy. In *Proceedings of the 17th Pacific Asia conference on language, information and computation* (pp. 152–158).

Hwang, W., Lee, D., Cho, K., Lee, H., & Seo, M. (2022). A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, *35*, 32537–32551.

Jiang, C., & Yang, X. (2023). Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the nineteenth international conference on artificial intelligence and law* (pp. 417–421).

Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the supreme court of the United States. *PLoS One*, *12*(4), Article e0174698.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT* (p. 2).

Kort, F. (1957). Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, *51*(1), 1–12.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Le, Y., Xiao, S., Xiao, Z., & Li, K. (2023). Topology-aware multi-task learning framework for civil case judgment prediction. *Expert Systems with Applications*, Article 122103.

Li, C., Ge, J., Cheng, K., Luo, B., & Chang, V. (2022). Statute recommendation: Re-ranking statutes by modeling case-statute relation with interpretable hand-crafted features. *Information Sciences*, *607*, 1023–1040.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).

Liu, D., Du, W., Li, L., Pan, W., & Ming, Z. (2022). Augmenting legal judgment prediction with contrastive case relations. In *Proceedings of the 29th international conference on computational linguistics* (pp. 2658–2667). Gyeongju, Republic of Korea: International Committee on Computational Linguistics, URL: https://aclanthology.org/2022.coling-1.235.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (2017). Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2727–2736). Copenhagen, Denmark: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D17-1289, URL: https://aclanthology.org/D17-1289.

Lyu, Y., Wang, Z., Ren, Z., Ren, P., Chen, Z., Liu, X., et al. (2022). Improving legal judgment prediction through reinforced criminal element extraction. *Information Processing & Management*, *59*(1), Article 102780.

Makridakis, S. (2017). The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46–60.

Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., Bhattacharya, A., et al. (2021). ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 4046–4062). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.313, URL: https://aclanthology.org/2021.acl-long.313.

Mamakas, D., Tsotsi, P., Androutsopoulos, I., & Chalkidis, I. (2022). Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. arXiv preprint arXiv:2211.00974.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., & Dehak, N. (2019). Hierarchical transformers for long document classification. In *2019 IEEE automatic speech recognition and understanding workshop* (pp. 838–844). IEEE.

Park, H., Vyas, Y., & Shah, K. (2022). Efficient classification of long documents using transformers. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 702–709). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-short.79, URL: https://aclanthology.org/2022.acl-short.79.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67, URL: http://jmlr.org/papers/v21/20-074.html.

Souza, F., Nogueira, R., & Lotufo, R. (2019). Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv:1909.10649.

Şulea, O.-M., Zampieri, M., Vela, M., & van Genabith, J. (2017). Predicting the law area and decisions of french supreme court cases. In *Proceedings of the international conference recent advances in natural language processing, RANLP 2017* (pp. 716–722). Varna, Bulgaria: INCOMA Ltd..

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wan, L., Papageorgiou, G., Seddon, M., & Bernardoni, M. (2019). Long-length legal document classification. arXiv preprint arXiv:1912.06905.

Wang, X., Han, X., Liu, Z., Sun, M., & Li, P. (2019). Adversarial training for weakly supervised event detection. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 998–1008). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1105, URL: https://aclanthology.org/N19-1105.

Wong, K.-F., Li, W., Xu, R., & Zhang, Z.-s. (2022). *Introduction to Chinese natural language processing*. Springer Nature.

Wu, Y., Liu, Y., Lu, W., Zhang, Y., Feng, J., Sun, C., et al. (2022). Towards interactivity and interpretability: A rationale-based legal judgment prediction framework. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 4787–4799). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.emnlp-main.316, URL: https://aclanthology.org/2022.emnlp-main.316.

Wu, Y., Zhou, S., Liu, Y., Lu, W., Liu, X., Zhang, Y., et al. (2023). Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration. In H. Bouamor, J. Pino, K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing*. Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.emnlp-main.740, URL: https://aclanthology.org/2023.emnlp-main.740.

Xiao, C., Hu, X., Liu, Z., Tu, C., & Sun, M. (2021). Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, *2*, 79–84.

Xiao, S., Liu, Z., Zhang, P., & Muennighof, N. (2023). C-pack: Packaged resources to advance general chinese embedding. arXiv preprint arXiv:2309.07597.

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 preprint arXiv:1807.02478.

Xu, N., Wang, P., Chen, L., Pan, L., Wang, X., & Zhao, J. (2020). Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3086–3095). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.280, URL: https://aclanthology.org/2020.acl-main.280.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., et al. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 483–498). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.41, URL: https://aclanthology.org/2021.naacl-main.41.

Yang, W., Jia, W., Zhou, X., & Luo, Y. (2019). Legal judgment prediction via multi-perspective bi-feedback network. arXiv preprint arXiv:1905.03969.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1480–1489).

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., et al. (2024). Qwen2 technical report. arXiv preprint arXiv:2407.10671.

Yao, F., Xiao, C., Wang, X., Liu, Z., Hou, L., Tu, C., et al. (2022). LEVEN: A large-scale Chinese legal event detection dataset. In *Findings of the association for computational linguistics: ACL 2022* (pp. 183–201). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.findings-acl.17, URL: https://aclanthology.org/2022.findings-acl.17.

Yue, L., Liu, Q., Jin, B., Wu, H., Zhang, K., An, Y., et al. (2021). 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* (pp. 973–982).

Yue, L., Liu, Q., Wu, H., An, Y., Wang, L., Yuan, S., et al. (2021). Circumstances enhanced criminal court view generation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 1855–1859).

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., et al. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, *33*, 17283–17297.

Zhang, Z., Han, X., Zhou, H., Ke, P., Gu, Y., Ye, D., et al. (2021). CPM: A large-scale generative Chinese pre-trained language model. *AI Open*, *2*, 93–99.

Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., & Sun, M. (2018). Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3540–3549). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1390, URL: https://aclanthology.org/D18-1390.

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.

Zhong, H., Zhang, Z., Liu, Z., & Sun, M. (2019). *Open Chinese language pre-trained model zoo*: *Technical report*, URL: https://github.com/thunlp/openclap.