

Beyond text: Fusing multi-modal legal knowledge for legal judgment prediction

Qihui Zhao^a, Tianhan Gao^b, Nan Guo^{c,*}

^a School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, Qinhuangdao, 066004, China

^b Software College, Northeastern University, Shenyang, 110169, China

^c School of Computer Science and Engineering, Northeastern University, Shenyang, 110169, China

ARTICLE INFO

Keywords:

Legal intelligence

Legal judgment prediction

Multi-modal legal knowledge fusion

Pre-trained language model

ABSTRACT

The objective of the Legal Judgment Prediction (LJP) task is to predict judgment outcomes based on the fact description texts within legal cases. Existing LJP methods are confined to leveraging knowledge inherent only within the dataset itself, often failing to achieve satisfactory performance when factual descriptions contain text prone to causing erroneous judgments. Consequently, extracting and utilizing external legal knowledge represents a critical challenge that the LJP task urgently needs to overcome. To address the aforementioned issues, this study proposes a legal judgment framework named MLK-LJP, which pioneers the integration of multi-granularity, multi-modal legal knowledge into the LJP task. MLK-LJP comprises two primary modules: Multimodal Legal Knowledge Extraction (MLKE) and Multi-modal Legal Knowledge Fusion (MLKF). Specifically: 1) In the MLKE module, we devise distinct methods to acquire five types of multi-modal legal knowledge: Legal article knowledge, legal event knowledge, legal relation knowledge, quantitative evidence knowledge, and image evidence knowledge. 2) In the MLKF module, we first design a Legal Knowledge Experts Fusion mechanism. This mechanism leverages a Graph Neural Network to capture collaborative signals among the five legal knowledge expert types. Subsequently, the fused multi-modal legal knowledge is allocated across different layers of a Transformer model. This legal knowledge enhanced Transformer model, combined with LJP prompts, is used to predict the LJP outcomes. Extensive experiments conducted on the three LJP datasets demonstrate the effectiveness and validity of MLK-LJP in comparison to state-of-the-art methods.

1. Introduction

The increasing digitization of legal processes worldwide has spurred significant interest in developing intelligent systems capable of assisting legal professionals and improving judicial efficiency. Among various burgeoning applications of Artificial Intelligence (AI) in the legal domain, Legal Judgment Prediction (LJP) has emerged as a critical research frontier. Tasked with automatically predicting judicial decisions based on textual descriptions of case facts, LJP holds immense potential for applications ranging from legal assistance and case outcome analysis to promoting judicial consistency. Recent advancements, particularly those leveraging sophisticated deep learning models like Transformers, have markedly improved the ability to capture complex semantic nuances within lengthy and intricate fact descriptions, forming the backbone of current state-of-the-art LJP approaches. Fig. 1 shows the example of LJP on the dataset CAIL2018 [1].

LJP has attracted great interest for several years. Early works focus on statistical machine learning on handcrafted manual features [2–4].

In recent years, deep learning has dominated the major advances in the natural language processing (NLP) field. Structures based on neural networks, such as convolutional neural network (CNN), recurrent neural network (RNN), graph neural networks (GNN) and Transformer, have been shown to have great success in learning text feature representations [5–8]. Pre-trained Language Models (PLMs) play a critical role in the successes of deep learning models, which significantly boosts the performance of a variety of NLP tasks. Many of the legal intelligence tasks have achieved impressive improvement, from similar case retrieval to legal text abstract [9].

However, despite these notable progresses, existing LJP methods predominantly operate within a constrained information space, relying solely on the knowledge intrinsically present within the provided fact descriptions. This inherent limitation renders them vulnerable when encountering factual narratives that are ambiguous, incomplete, or contain misleading details, often leading to unsatisfactory or even erroneous predictions. The crux of the issue lies in the neglect of vast amounts of external legal knowledge—such as relevant legal statutes, precedents

* Corresponding author.

E-mail addresses: zhaoqihui@neuq.edu.cn (Q. Zhao), gaoth@mail.neu.edu.cn (T. Gao), guonan@mail.neu.edu.cn (N. Guo).

<https://doi.org/10.1016/j.knosys.2025.114358>

Received 28 April 2025; Received in revised form 8 August 2025; Accepted 27 August 2025

Available online 31 August 2025

0950-7051/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

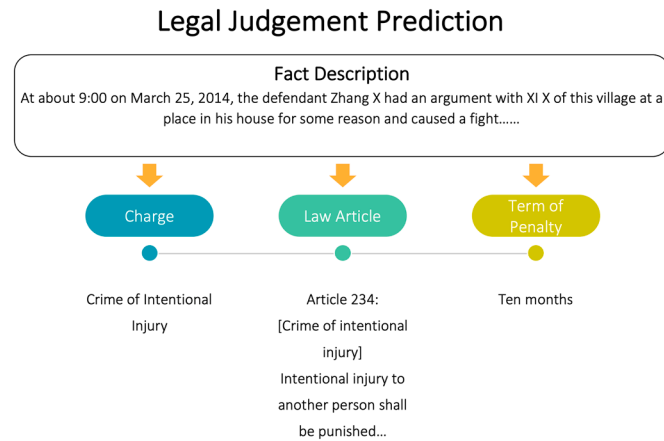


Fig. 1. An example of LJP datasets. LJP is aim to predict law article, charge, and term of penalty based on the fact description.

from similar cases, established legal principles, and even non-textual evidence—which human judges routinely employ. Consequently, a critical and largely unmet challenge within the LJP field is twofold: first, identifying and acquiring the pertinent external legal knowledge relevant to a given case, encompassing potentially diverse modalities; and second, developing effective mechanisms to seamlessly fuse this heterogeneous knowledge with the factual context to inform the final prediction. Addressing this gap is paramount for pushing LJP systems towards greater robustness and real-world applicability.

To bridge this critical gap and empower LJP with deeper legal understanding, this paper argues for the necessity of incorporating external, multi-faceted legal intelligence. Motivated by this, we introduce MLK-LJP (Multi-modal Legal Knowledge for Legal Judgment Prediction), a novel judicial judgment framework specifically designed to tackle the aforementioned challenges. MLK-LJP pioneers the systematic integration of multi-granularity, multi-modal external legal knowledge into the LJP task. At its core, MLK-LJP comprises two synergistic modules: (1) a Multimodal Legal Knowledge Extraction (MLKE) module, which devises distinct strategies to acquire five crucial types of legal knowledge—namely, legal articles, legal events, legal relations, quantitative evidence, and significantly, image-based evidence—thereby addressing the “what knowledge to acquire” challenge with unprecedented breadth. (2) A Multi-modal Legal Knowledge Fusion (MLKF) module, which tackles the “how to fuse” challenge by employing a Graph Neural Network (GNN) to explicitly model the collaborative signals and interdependencies between these diverse knowledge ‘experts’, subsequently allocating the fused, enriched knowledge representation across different layers of a Transformer architecture for enhanced judgment prediction, guided by tailored LJP prompts.

The characteristics and contributions of MLK-LJP are summarized as:

- To break through the bottleneck of only using uni-modal data in existing methods, we for the first time interact legal judgment prediction in a multi-modal legal knowledge mixture-of-experts framework through neural networks. Subsequently, the proposed method is applicable to extract fine-grained legal features, resulting in improved performance of the task;
- For multi-granularity and multi-modal legal knowledge, we design different methods to acquire knowledge of law articles, knowledge of legal events, knowledge of legal relations, numerical evidence, and multi-modal (image) evidence, respectively. All these knowledge are taken on the existing datasets to obtain without duplicating the work.
- For legal knowledge fusion, we aim to not only preserve the effective information in the original features, but also to enhance the importance of its features in the fusion of text encoders to fuse legal knowl-

edge in a more detailed way. For these purposes, we design a Legal knowledge experts fusion and introduce legal knowledge-enhanced mechanism in text encoder.

- Extensive experiments conducted on three benchmark datasets, demonstrate the substantial superiority of MLK-LJP, achieving new state-of-the-art performance across virtually almost standard evaluation metrics and validating the effectiveness of incorporating multi-modal legal knowledge.

The remainder of the paper is organised as follows. Related work is reviewed in Section 2. Section 3 outlines our method. Section 4 presents the evaluation results. Section 5 depicts the analysis of results. We conclude our work in Section 6.

2. Related work

2.1. Legal judgment prediction

Legal Judgment Prediction (LJP), a pivotal task within the broader field of legal intelligence, typically involves predicting judicial outcomes such as applicable law articles, charges, and prison terms based on textual fact descriptions. It is often framed as a specialized text classification problem. Existing approaches broadly fall into two main categories: traditional methods relying on statistical or feature-engineering techniques, and more recent methods leveraging deep neural networks.

2.1.1. Traditional approaches

Early explorations into LJP involved quantitative analysis. Pioneering work by Kort [2] applied statistical methods to predict court decisions, although these early studies were often limited to small datasets with few outcome labels. Subsequently, traditional machine learning techniques were employed. For instance, Lin et al. [3] utilized Conditional Random Fields (CRF) with manually defined legal element labels for charge prediction. Katz et al. [4] developed a time-evolving random forest classifier, relying on specific feature engineering to predict case outcomes. A significant limitation of these machine learning-based approaches is their dependence on extensive manual feature engineering, a process that is both labor-intensive and potentially domain-specific.

2.1.2. Deep learning-based approaches

With the advent of deep learning and its remarkable success in Natural Language Processing (NLP), LJP research has increasingly adopted neural network architectures [5–8]. These deep learning approaches can be further categorized based on their primary focus: designing novel model architectures tailored for LJP or explicitly incorporating external legal knowledge. **Novel Model Architectures:** A significant body of work focuses on developing specialized neural architectures to better capture the nuances of legal text. Early examples include combining FastText with TextCNN [10]. Hierarchical models, often employing attention mechanisms, were proposed to predict charges by modeling the structure of legal documents [11]. Recursive attention networks were introduced by Yang et al. [12] to map fact descriptions to relevant law articles. Recognizing the sequential nature of judicial decision-making, Ma et al. [13] designed a multi-task learning method simulating the judge’s process by integrating plaintiff claims and court debate data. Specific sub-tasks also received focused architectural attention; for instance, Chen et al. [14] utilized gating mechanisms to improve penalty term prediction, while Pan et al. [15] employed multi-level attention to handle cases with multiple defendants. More recently, Graph Neural Networks (GNNs) have been explored, with Zhao et al. [16] building and fusing multiple graphs for prediction. Framework-level innovations include topology-aware multi-task learning [17] and recall-and-ranking frameworks that decompose LJP into candidate generation and verification stages [18]. Other approaches transform LJP into a node classification problem using graph reasoning and contrastive learning [19]. **Integration of Legal Knowledge:** Another line of research focuses on

enhancing LJP models by explicitly incorporating external legal knowledge, primarily textual. Kang et al. [20] improved charge prediction by using accusation definitions to refine fact description representations. Luo et al. [21] integrated information from relevant law articles into an attention-based RNN model. To address data sparsity issues, Hu et al. [22] summarized legal attributes and applied them specifically to improve predictions for infrequent charges. Yue et al. [23] proposed utilizing intermediate sub-task results (like identifying distinct factual circumstances) to inform other predictions and explicitly modeled potential confusion between similar charges or articles. Sun et al. [24] explored prompt learning, using an external knowledge base to extract keywords from facts and incorporating them into prompt templates to guide Pre-trained Language Models (PLMs).

While the aforementioned approaches have significantly advanced the LJP field, particularly those incorporating external textual knowledge, they largely overlook the potential of leveraging diverse, multi-modal legal knowledge sources (such as quantitative data or visual evidence). Our work addresses this gap by proposing MLK-LJP, a framework specifically designed to systematically extract, fuse, and utilize multi-modal external legal knowledge to achieve more accurate and robust legal judgment predictions.

2.2. Prompt learning

Prompt learning represents a significant paradigm shift in NLP, diverging from traditional fine-tuning [25]. Fine-tuning typically adapts PLMs to downstream tasks using distinct, task-specific objective functions, which may not align optimally with the model's original pre-training methodology. In contrast, prompt learning reformulates downstream tasks to mirror the formats PLMs encountered during pre-training, aiming to more effectively elicit the knowledge embedded within them. This approach has demonstrated success across various NLP domains. Research in prompt learning has explored several avenues. For instance, Gu et al. [26] investigated soft prompts for pre-

training large models and applied them to few-shot NLU tasks. Liu et al. [27] implemented prefix-tuning, adding continuous prompts to each Transformer layer and utilizing the [CLS] token for prediction. Addressing few-shot NLU, Wang et al. [28] introduced an auxiliary self-supervised task to improve PLM adaptation to the prompt format. Methodological refinements include Hu et al.'s [29] use of answer engineering to inject knowledge for text classification, and Cui et al.'s [30] application of contrastive learning to explicitly learn soft labels, enhancing few-shot classification performance. Prompt learning's application extends beyond standard NLU. Rao et al. [31] designed instance-level prompts by integrating visual features, creating data-specific prompts. Cho et al. [32] proposed a unified framework for multimodal tasks, employing conditional text generation (e.g., generating labels from image-text pairs). In sequence labeling, Cui et al. [33] pioneered prompt application by constructing N-gram candidates and using BART with templates for entity classification. Lee et al. [34] further enhanced sequence labeling by integrating prompt template information into existing models. Inspired by the demonstrated versatility and effectiveness of prompt learning in NLP, this work investigates its application to the task of legal judgment prediction.

3. Our method

3.1. Problem formulation

We formulate the Legal Judgment Prediction (LJP) task by defining its core components:

- **Fact Description (F):** The textual narrative of a case from a legal document, detailing the pertinent facts, events, and actions.
- **Law Articles (D_l):** The set of legal statutes or provisions from the relevant legal code (e.g., Criminal Code) applicable to the case facts F .
- **Charges (D_c):** The set of formal criminal offenses (e.g., theft, fraud) determined from the case facts F and relevant law articles D_l .

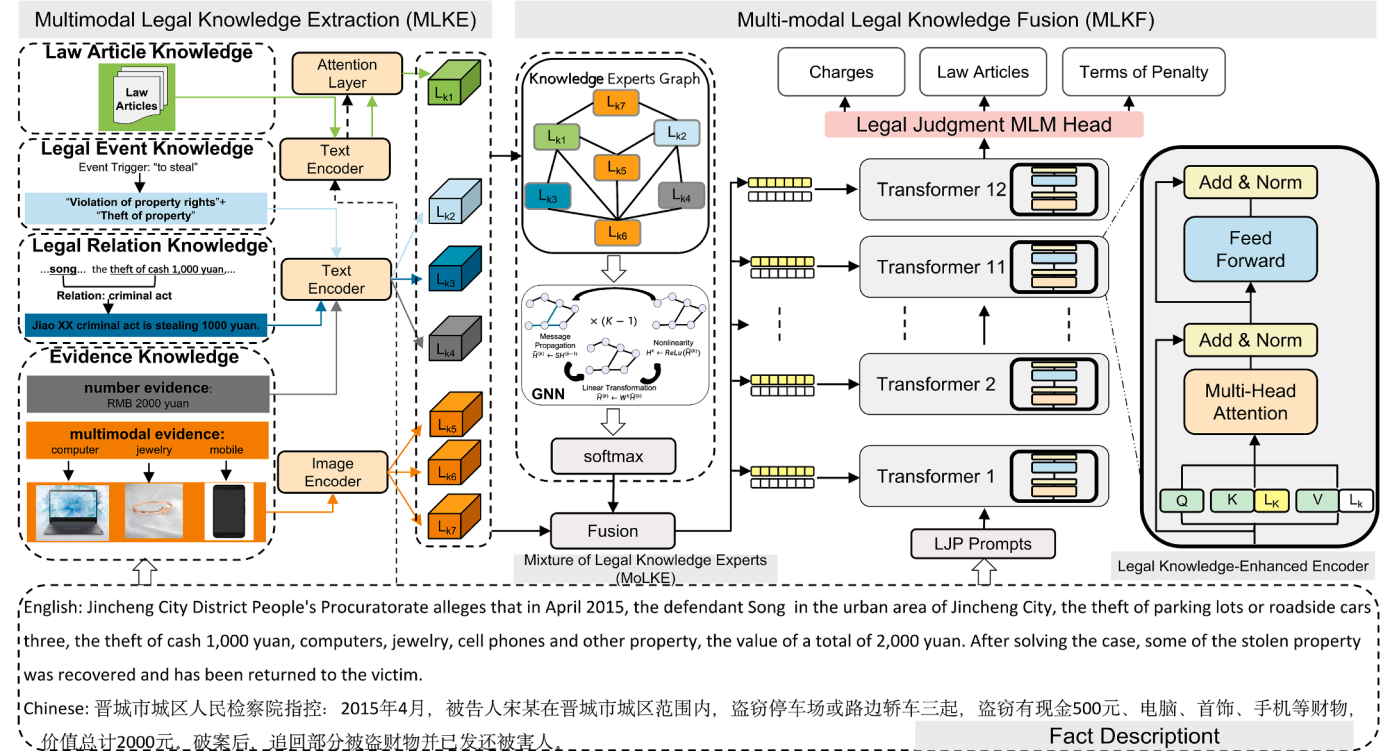


Fig. 2. The proposed LJP architecture: MLK-LJP. MLK-LJP first acquires multiple legal knowledge related to fact descriptions through multi-modal legal knowledge extraction, then processes the knowledge into dimension-specific vectors through multi-modal legal knowledge fusion.

- **Term of Penalty (D_t):** The duration of imprisonment determined for the charges D_c , based on the facts F and articles D_l .
- **Legal Judgment Prediction (LJP):** The LJP task aims to automatically predict a case's judicial outcome from its fact description F . We formulate this as a multi-task learning problem where a model M , given the input F , must simultaneously predict three key judgment elements: the applicable charges (D_c), relevant law articles (D_l), and the term of penalty (D_t). As illustrated in Fig. 1, the overall function is formalized as: $M(F) = \{D_c, D_l, D_t\}$.

3.2. Method overview

As shown in Fig. 2, Our method has two components: multi-modal legal knowledge extraction (MLKE) and Multi-modal Legal Knowledge Fusion (MLKF). The input of our proposed method includes the fact description which consists of x_1, x_2, \dots, x_n and all law articles which contains l_1, l_2, \dots, l_m . MLK-LJP is handled in two branches after input. The first branch is legal knowledge extraction and multi-modal legal knowledge fusion. The second branch is input to the legal knowledge enhanced pre-trained model by designed LJP prompts.

3.3. Multi-modal legal knowledge extraction (MLKE)

The MLKE component aims to extract and integrate diverse legal knowledge crucial for the LJP task. This knowledge encompasses two modalities: textual information (including relevant law articles, legal relations, legal events, and quantitative evidence) and visual information (image evidence). This subsection details the extraction process for law articles knowledge.

3.3.1. Law articles knowledge

We leverage an attention mechanism to derive knowledge representations from relevant law articles based on the fact description. First, the input fact description, represented as a sequence of tokens $S = \{x_1, x_2, \dots, x_l\}$, is processed by a text encoder:

$$\mathbf{H}_{fact} = \mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_l = \text{Encoder}(x_1, x_2, \dots, x_l) \quad (1)$$

where \mathbf{H}_{fact} contains the hidden state vectors corresponding to each token in the fact description, and l is the sequence length. A summary representation of the fact description, the context vector $\hat{\mathbf{h}}$, is obtained via max-pooling over these hidden states:

$$\hat{\mathbf{h}} = \text{maxpooling}(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_l) \quad (2)$$

Finally, an attention mechanism computes the relevance between the fact context $\hat{\mathbf{h}}$ and each law article context \mathbf{c}_j . The attention score α_j for the j th law article is calculated as:

$$\alpha_j = \hat{\mathbf{h}} \mathbf{W}_1 \mathbf{c}_j$$

$$K_{LA} = \sum \frac{\exp(\alpha_j)}{\sum_{k=1} \exp(\alpha_k)} \quad (3)$$

where K_{LA} represents the aggregated knowledge from law articles, weighted by their relevance to the input fact description. \mathbf{W}_1 is a trainable matrix.

3.3.2. Legal events knowledge

We extract legal event knowledge using a model trained on the LEVEN dataset [35], whose schema categorizes events as shown in Table 1. The trained model detects event triggers (e.g., "to steal") in the LJP fact descriptions and generates a structured knowledge representation, such as the triplet "Prohibited acts-Property infringement-Theft". This fine-grained event knowledge, denoted K_{ED} , is obtained by feeding the fact description into the pre-trained DMBERT model [35].

3.3.3. Legal relation knowledge

Inspired by [36], we use prompt-based learning with Lawformer [37] to extract predefined legal relation triplets: (Head Entity, Relation, Tail Entity). We target three crucial relations: Criminal Purpose, Criminal Act, and Post-Criminal State, for which we manually annotated 1000 legal documents. In each triplet, the Head Entity is the defendant's name, while the Tail Entity is the textual span describing the relation's object (e.g., motive or action). For each extracted triplet, we construct a descriptive sentence (e.g., "Han X criminal purpose profit oriented" from Table 2) and encode it with Lawformer to produce the knowledge vector K_{RE} :

$$K_{RE} = E(S) \quad (4)$$

3.3.4. Quantitative evidence knowledge

The amount of the offense in the fact description is important evidence for predicting the terms of penalties in certain types of cases (e.g., financial cases). However, the amount of crime in factual descriptions is randomly distributed. As a result, it is difficult for the model to directly derive the exact total amount of the crime and to predict the correct sentence based on scattered numbers. Specifically, we train the quantitative evidence Knowledge model (BERT-CRF NER model) on the dataset CAE-NER. Then, the sum of the numbers is regarded as the final crime amount. Assuming that the model recognizes the Quantitative evidence as 3000 yuan, the method generates the sentence, "The total amount involved in the case is 3000 yuan" which is input in a text encoder and noted as K_{QE} .

3.3.5. Image evidence knowledge

Object evidence is generally present in fact descriptions and these items play a very important role in convictions. Our method considers the use of visual modality (images of object) to enhance the representation of objects evidence. First the method uses named entity recognition model trained on E-Commercial NER Dataset to identify the objects evidence in the text (5 objects for each instance), then 5 images that are searched in the Bing search engine. We use CLIP to calculate the similarity and filter the 3 images with the highest match scores out of the 25 images as the images evidence.

We use two level of image evidence to obtain diverse image representations, namely full image evidence and object evidence. For the full image encoder, we input the image into CLIP's vision transformer to get the pixel-level full image representation K_{IF} . For the object detection encoder, we first input the image into the Faster-RCNN [38] to obtain the top 3 significant regions and feed them into the regions encoding module, CLIP's vision transformer, to obtain the pixel-level representation. Finally, we concatenate the representations of the 3 object images together to get the final pixel-level image objects representation K_{IO} .

Table 1

Legal event over the top-level event types and the corresponding categories and samples.

Top-type event	Category	N of type	Sub-type examples	Bottom-type examples
General_behaviors	Behavior	40	Property_infringement, Organizing	Selling, Employing, Manufacturing
Prohibited_acts	Behavior	40	Civil_activities, Complicity	Killing, Blackmail, Theft
Judicature_related	Behavior	13	None	Arrest, Surrendering
Consequences	Result	7	None	Death, Injury, Being_trapped
Accident	Result	4	None	Traffic_accident, Fire_accident
Natural_disaster	Majeure	4	None	Drought, Flood_and_waterlogging

Table 2
Examples of three types of legal relations.

Head entity	Relation	Tail entity
Han X	criminal purpose	profit oriented
Zhang X	criminal purpose	illegal possession
Chen X	criminal act	steal mobile phone
Fu X	criminal act	punches to the face
Xiang X	post-criminal state	poor guilty plea
Huang X	post-criminal state	good guilty plea

3.4. Multi-modal legal knowledge fusion (MLKF)

3.4.1. Mixture of legal knowledge experts (MoLKE)

To effectively fuse the diverse extracted legal knowledge representations (K_{LA} , K_{RE} , K_{ED} , K_{QE} , K_{IO_1} , K_{IO_2} , K_{IO_3}), we propose a Mixture of Legal Knowledge Experts (MoLKE) architecture. Traditional Mixture-of-Experts (MoE) models often use a simple gating function (e.g., Softmax router) to assign input tokens to experts, which can lead to suboptimal collaboration between experts and potential load imbalance issues.

Graph-enhanced router. To mitigate these limitations, we introduce a novel graph-based router. This router explicitly models the interactions between different knowledge sources (experts) when determining how to weight them for a given input context. We construct an MoE graph $G = (V, E)$, where the node set V includes nodes representing each of the N legal knowledge experts (e_1, e_2, \dots, e_N) and potentially nodes representing the input context. Instead of a simple Softmax, we employ a Graph Neural Network (GNN) within the router. For a given input representation x and a specific expert node e_i , the GNN aggregates information from e_i 's neighbors $N(e_i)$ in the graph. These neighbors can include other expert nodes and the input node x itself, allowing the GNN to capture collaborative information. The GNN learns an interaction-aware representation for each expert conditioned on the input and other experts. This information signifies not only the learning capability of each expert concerning the input token node but also captures the collaborative information among all experts. The specific formulation is conceptualized as follows:

$$\mathcal{R}_{GNN}(x)_i = \mathcal{R}(F(GNN(e_{ki}, N(e_{ki})))) \quad (5)$$

where $GNN(-)$ represents the Graph Neural Network, which learns the representation for an expert node e_{ki} by utilizing information from its neighbors $N(e_{ki})$, including other experts and the input token node x . $\mathcal{R}(-)$ is the Softmax function used to assign probability weights across all experts. Following the sparse gating strategy, we utilize only the Top-K experts during the feedforward process.

Knowledge integration into transformer layers. While the knowledge vectors K_i are derived from Transformer-based models, effectively integrating this fused knowledge or the individual weighted knowledge vectors into different layers of the main LJP Transformer model requires careful consideration. We explore two methods: a standard Multi-Layer Perceptron (MLP) mapping and a dedicated Knowledge Allocation Module (KAM).

The KMM aims to dynamically determine how much influence each type of legal knowledge should have at each specific layer of the main Transformer. It calculates a layer-specific mapping ratio vector $ar^l \in \mathbb{R}_N$ (where $N = 7$ is the number of knowledge types in our case). Each element ar^l_i represents the relevance or weight assigned to the i th knowledge type K_i for layer l . This ratio is computed based on an aggregated representation of the activated knowledge:

$$K_{agg}(x) = \frac{1}{|\text{TopK}|} \sum_{i \in \text{TopK}} g_i(x) K_i \quad (6)$$

$$ar^l = \text{Softmax}(\text{RReLU}(W_2 K_{agg}(x))) \quad (7)$$

Where W_2 is a trainable weight matrix specific to layer l that projects the aggregated knowledge vector $K_{agg}(x)$ into N dimensions. RReLU (Randomized Leaky ReLU) is used as the activation function, and Softmax normalizes these scores into a probability distribution $ar^l = (ar^l_1, ar^l_2, \dots, ar^l_N)$ ensuring $\sum_i ar^l_i = 1$.

Using these layer-specific ratios, we compute the final knowledge vector V_k^l for the i th knowledge type to be injected into layer l :

$$V_k^l = ar^l K \quad (8)$$

This scales each original knowledge vector K_i based on its dynamically calculated relevance ar^l_i for the specific layer l .

Finally, the scaled knowledge vectors for all N types pertinent to layer l are concatenated to form the complete knowledge representation \hat{V}_k^l for that layer:

$$\hat{V}_k^l = [V_{k_{LA}}^l, \dots, V_{k_{IO}}^l] \quad (9)$$

This concatenated vector \hat{V}_k^l is then integrated into the computations of the l th layer of the main Transformer model.

3.4.2. Legal knowledge-enhanced text encoder (LKTE)

To infuse the extracted and fused legal knowledge into the contextual text representations, we enhance the self-attention mechanism within each layer of the text encoder. Inspired by prefix-tuning approaches, we treat the layer-specific legal knowledge representation \hat{V}_k^l as a dynamic “prefix” that informs the attention calculation. This allows the model to update text representations while simultaneously considering the relevant legal knowledge context at each processing stage. Specifically, at each Transformer layer l , the standard self-attention mechanism computes Q^l (query), K^l (key), and V^l (value) matrices from the layer's input sequence embeddings. To incorporate the legal knowledge \hat{V}_k^l , we first project it into the Key and Value spaces using layer-specific linear transformations:

$$\hat{L}_k^l = \hat{V}_k^l W_l^k \quad (10)$$

$$\hat{L}_v^l = \hat{V}_k^l W_l^v \quad (11)$$

where \hat{V}_k^l is the concatenated knowledge vector for layer l , and W_l^k, W_l^v are trainable weight matrices for layer l . These projections map the fused knowledge into vectors \hat{L}_k^l and \hat{L}_v^l which have dimensions compatible with the text's Key and Value matrices, respectively. These vectors represent the legal knowledge contextualized for the key-query matching and value retrieval parts of the attention mechanism.

Note that the Query matrix Q^l remains unchanged. The self-attention scores are then computed using the original text queries attending to the extended keys, and the output is formed using the extended values:

$$K_Attention^l = \text{softmax}\left(\frac{Q^l [\hat{L}_k^l, K^l]^T}{\sqrt{d}}\right) [\hat{L}_v^l, V^l] \quad (12)$$

The output of this knowledge-enhanced attention mechanism replaces the output of the standard self-attention sub-layer within the l th Transformer encoder layer. By performing this integration at each layer, the final textual representations encode rich semantic information derived from both the input fact description and the dynamically weighted, relevant legal knowledge, ultimately benefiting the LJP task.

3.4.3. Legal judgment prediction by prompt learning

The prediction module uses prompt learning to reframe the LJP task.

Prompt template construction. We use a manually crafted prefix-prompt template that precedes the input fact description X :

$$\begin{aligned} pt = & [[CLS], p_1, p_2, \dots, p_{12}, [MASK], p_{13}, p_{14}, \\ & [MASK], [MASK], [MASK], [MASK], [MASK], \\ & \dots, p_m, [MASK], [SEP], X, [SEP]] \end{aligned} \quad (13)$$

The template includes task-specific instructions and [MASK] tokens as placeholders for the prediction targets (articles, charges, terms).

Verbalizer construction. The verbalizer maps the model’s vocabulary-level predictions at each [MASK] position to the structured LJP label space. For each [MASK] position, we take the argmax of the logits to get the predicted token:

$$\text{predict}[n] = \text{argmax}_{\text{logits}}[\text{MASK}_n] \quad (14)$$

where n represents each [MASK] position, $[\text{MASK}_n]$ is all positions to be predicted, and predict is the set of prediction results for all [MASK] positions in the input text. $\text{logits}[\text{MASK}_n]$ represents a one-dimensional vector which is the language model vocabulary probability of all words in the vocabulary at the predicted n th position. This token is then matched to the most similar candidate label in the corresponding label space using Edit Distance (for articles and terms) or Jaro Similarity (for charges).

Prediction and learning objective. The joint probability of predicting the correct set of labels y is the product of the probabilities at each corresponding [MASK] position:

$$p(y | x) = \prod_{j=1}^n p(\text{MASK}_j = \phi_j(y) | pt) \quad (15)$$

where n is the total number of masked positions, y_j is the specific label corresponding to the j th, and mask and $\phi_j(y_j)$ maps that label to its verbalizer token(s). The final learning objective of our work is to maximize the average log-likelihood of the correct label predictions across the entire training corpus \mathcal{X} :

$$V = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \log \prod_{j=1}^n p(\text{MASK}_j = \phi_j(y) | pt) \quad (16)$$

4. Experiments and results

4.1. Datasets

Our experiments are carried out on public datasets named Chinese AI and Law challenge (CAIL2018) which is made up of two datasets (CAIL-small and CAIL-big) [1]. Criminal cases published by the Supreme People’s Court can be found in CAIL2018. In each case, there are two parts: the description of the facts and the corresponding judgment outcomes. In order to process the data, referring to pipeline of LADAN [39], we first filter out infrequent charges and law articles and retain only those with frequencies greater than 100. Then the term of penalty is split into 11 non-overlapping intervals. Besides, our model aims to be consistent with state-of-the-art methods, we filter out samples with multi-labels and less than ten words. Table 3 shows the detailed statistics of the datasets. It can be seen that the number of samples in the Training set is 101,685 and the number of samples in the Test set is 26,766 in the CAIL-small. CAIL-small contains 103 types of law articles, 119 types of charges, and 11 types of terms. In the CAIL-big, the number of samples in the training set is 1,588,894 and the number of samples in the test set is 185,228. The CAIL-big contains 118 kinds of laws, 130 kinds of charges, and 11 kinds of terms. The inputs of LAIC-2021 are the same as CAIL2018, and since this paper does not deal with multi-law articles prediction, here we remove the law articles prediction subtask and only compare the charge prediction and term of penalty prediction.

Table 3
Dataset details.

Dataset	CAIL-small	CAIL-big	LAIC-2021
Training Set Cases	101,685	1,588,894	79169
Test Set Cases	26,766	185,228	9896
Law Articles	103	118	–
Charges	119	130	42
Term of Penalty	11	11	9

4.2. Evaluation metrics

Following previous works [21,39–41], we introduce accuracy (Acc.), macro-precision (MP), macro-recall (MR) and macro-F1 score(F1). The calculation formulas of the four evaluations are shown below:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{Macro} - P &= \frac{1}{N} \sum_{i=1}^N \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \\ \text{Macro} - R &= \frac{1}{N} \sum_{i=1}^N \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \\ \text{Macro} - F1 &= \frac{1}{N} \sum_{i=1}^N \frac{2(\text{Precision}_i * \text{Recall}_i)}{\text{Precision}_i + \text{Recall}_i} \end{aligned} \quad (17)$$

where TP_i means that the predicted label is i , and the ground truth is i . TN_i indicates that the predicted label is not i , and the ground truth is not i . FP_i refers that the predicted label is i , and the ground truth is not i . FN_i represents that the predicted label is not i , and the ground truth is i .

4.3. Baseline

We select the following baseline models for comparison, including classic text classification baselines and state-of-the-art methods in the LJP task:

- **TFIDF + SVM** [42]. A traditional method using TF-IDF for text feature extraction and a Support Vector Machine (SVM) for classification.
- **CNN** [43]. Employs Convolutional Neural Networks (CNN) with multiple filter sizes to extract features from text, followed by a Softmax classifier.
- **RCNN** [44]. Combines Recurrent Neural Networks (RNN) and CNNs to capture both sequential context and local features for text classification.
- **HARNN** [45]. Utilizes RNNs with a Hierarchical Attention mechanism to model document structure and extract features, classified using Softmax.
- **BERT** [46]. Uses BERT to capture text features for the LJP task.
- **FLA** [21]. Integrates relevant law article information into the prediction process using attention mechanisms during fact description feature extraction.
- **TOPJUDGE** [40]. Proposes a multi-task learning framework that explicitly models dependencies between LJP subtasks using a directed acyclic graph.
- **MPBFN-WCA** [41]. Introduces a multi-task learning framework featuring multi-perspective forward prediction and backward verification mechanisms for LJP.
- **LADAN** [39]. Employs a Graph Neural Network (GNN) specifically designed to learn discriminative representations for legally similar but distinct law articles, aiding charge prediction.
- **NeurJudge** [23]. Proposes a framework that explicitly incorporates extracted “circumstances of crime” information to improve judgment prediction accuracy.
- **R²** [18]. Formulates LJP as a two-stage problem, first recalling a set of candidate results and then ranking them for the final prediction.
- **GraSCL** [19]. Transforms the text classification aspect of LJP into a node classification task on a graph structure, leveraging graph reasoning and supervised contrastive learning.

4.4. Experimental settings

We use PyTorch to implement the model. The optimization algorithm is Adam [47] with the learning rate set to 0.01 and the dropout set to 0.5. We use word segmentation tool named THULAC

Table 4
LJP results on CAIL-big.

	Law articles				Charges				Term of penalty			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
TFIDF + SVM	89.93	68.56	60.58	61.25	85.81	69.76	61.92	63.51	54.13	39.15	37.62	39.14
FLA	93.22	72.81	64.27	66.57	92.48	76.21	68.12	69.97	57.66	43.01	38.89	41.63
CNN	93.79	80.79	73.15	73.12	92.23	83.27	75.13	78.12	52.41	42.23	35.73	36.76
RCNN	93.98	80.93	73.26	73.13	93.50	85.89	77.03	79.65	53.62	43.43	36.18	37.67
HARNN	94.01	80.99	73.58	75.08	93.62	85.93	77.27	79.79	54.12	42.61	37.57	38.59
TOPJUDGE	95.83	82.10	71.94	74.32	95.77	85.95	77.11	79.58	58.09	47.73	42.47	44.07
MBPFN	95.67	84.00	74.40	76.44	94.37	85.60	75.86	77.98	55.48	47.27	38.26	40.01
LADAN	95.78	84.93	75.88	78.79	94.58	85.52	77.36	80.04	56.34	47.76	40.48	42.02
NeurJudge	95.59	84.01	75.54	77.06	94.12	85.48	77.21	79.83	55.52	47.25	40.76	42.03
R ²	97.24	87.44	82.89	84.46	97.23	90.77	86.67	88.23	61.27	54.32	47.99	50.10
GraSCL	97.92	90.84	87.79	88.84	97.88	93.17	90.73	91.62	65.60	57.89	56.69	57.06
MLK-LJP	97.51	91.74	87.23	88.99	97.92	93.25	90.77	91.74	68.14	58.64	57.27	58.12

(<https://github.com/thunlp/THULAC-Py-thon>), which is excellent in tackling Chinese. We terminate training if the validation loss does not drop for 100 consecutive epochs. Other parameter searches are carried out by searching for the dropout rate in {0.3, 0.5, 0.7}, and the batch size in {8, 16, 32, 64}. From a code implementation perspective, we employ scikit-learn (<https://scikit-learn.org/stable/>) for metric measurement. Lastly, all reported results are the mean values from three independent runs with different random initializations. For the baseline models in the experiments, we reproduce results of CNN and HARNN based on the open-source code and other results are obtained by running the original code provided in the paper.

4.5. Experimental results and discussion

This section first analyzes the performance of MLK-LJP on the three sub-tasks of CAIL-2018 datasets: relevant article prediction, charge prediction, and term prediction, with detailed results presented in Tables 4 and 5.

For relevant articles prediction, MLK-LJP demonstrates highly competitive, state-of-the-art performance. Specifically, on the CAIL-big dataset, it surpasses the previous best-performing models by achieving absolute gains of 0.9 % in Precision and 0.15 % in F1-score. On the CAIL-small dataset, MLK-LJP achieves even more significant improvements, boosting Precision by 0.09 %, Recall by 0.4 %, and F1-score by a notable 1.5 %. Consistent with prior findings, specialized LJP models (including TopJudge, MBPFN, LADAN, NeurJudge, R², and GraSCL) generally outperform standard text classification baselines, highlighting the value of architectures tailored to the legal domain. The strong performance of MLK-LJP can be attributed to its core design philosophy: the explicit integration of diverse legal knowledge sources (legal articles, events, relations, and multi-modal evidence). This approach appears to align well with the actual reasoning process in legal judgment and provides demonstrable benefits for the LJP task. When comparing MLK-LJP

specifically against LADAN [39] and NeurJudge [23] - models that also focus on incorporating legal article knowledge - several advantages of our approach emerge: (1) Efficient Knowledge Fusion: Unlike LADAN, which may rely more heavily on GNNs for article integration, MLK-LJP directly fuses article information into each self-attention layer of the Transformer. This strategy not only achieves superior performance but also potentially reduces the computational complexity often associated with GNN-based fusion across the entire dataset. (2) Handling Sparsity: MLK-LJP significantly outperforms both LADAN and NeurJudge on both datasets, suggesting its enhanced capability in handling less frequent or “long-tail” law articles, possibly due to the richer context provided by the multi-modal knowledge fusion. (3) Effective Guidance: The direct comparison validates that the specific mechanism employed by MLK-LJP for merging legal knowledge within the Transformer architecture provides highly effective guidance for the prediction tasks. Interestingly, we observe that GraSCL exhibits slightly better performance than MLK-LJP specifically on the relevant article recall metric for CAIL-big (by 0.56 %) and overall accuracy on CAIL-small (by 0.41 %). This nuanced result is likely attributable to GraSCL’s specialized design, which employs contrastive learning explicitly optimized for discriminating between relevant law articles. However, MLK-LJP’s overall superior performance across a broader range of metrics and tasks underscores the general effectiveness and robustness of its multi-modal knowledge integration strategy.

Regarding Charge Prediction, the objective is to determine the specific criminal charges based on the fact description. MLK-LJP integrates knowledge pertaining to legal articles, events, relations, quantitative data, and image evidence directly into the Transformer’s self-attention mechanisms. For charges lacking an explicitly linked law article in our knowledge extraction phase, we utilize the charge’s full name representation. As detailed in Tables 4 and 5, our method consistently outperforms all competing approaches on both datasets. Specifically, on CAIL-big, MLK-LJP achieves absolute gains of 0.14 % in

Table 5
LJP results on CAIL-small.

	Law articles				Charges				Term of penalty			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
TFIDF + SVM	76.52	43.21	40.12	39.68	79.81	45.86	42.72	42.77	33.32	27.66	24.99	24.64
FLA	77.72	75.21	74.12	72.78	80.98	79.11	77.92	76.77	36.32	30.81	28.22	27.83
CNN	78.61	75.86	74.60	73.59	82.23	81.57	79.73	78.82	35.20	32.96	29.09	29.68
RCNN	79.12	76.58	75.13	74.15	82.50	81.89	79.72	79.05	35.52	33.76	30.41	30.27
HARNN	79.73	75.05	76.54	74.67	83.41	82.23	82.27	80.79	35.95	34.50	31.04	31.18
TOPJUDGE	79.79	79.52	73.39	73.33	82.03	83.14	79.33	79.03	36.05	34.54	32.49	29.19
MPBFN	79.12	76.30	76.02	74.78	82.14	82.28	80.72	80.72	36.02	31.94	28.60	29.85
LADAN	82.34	78.79	77.59	76.80	84.83	83.33	82.80	82.85	39.35	36.94	33.25	34.05
NeurJudge	82.57	79.62	76.02	84.63	83.92	82.76	81.85	81.96	39.65	38.87	36.12	37.17
R ²	83.85	83.11	83.48	82.04	89.31	87.48	88.22	87.61	41.62	40.89	37.01	38.32
GraSCL	86.48	84.40	83.57	83.30	90.03	89.33	89.23	89.00	43.98	42.78	40.78	41.43
MLK-LJP	86.98	84.49	83.97	86.13	90.14	90.82	89.72	89.94	45.01	43.97	42.63	42.82

Table 6
LJP results on LAIC-2021.

	Charges				Term of penalty			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1
CNN	96.41	91.91	90.17	91.03	34.14	33.19	32.69	32.94
BERT	96.72	93.25	88.82	90.98	41.83	41.88	41.43	41.65
HARNN	95.11	88.31	84.93	86.59	33.43	32.07	30.44	31.23
TOPJUDGE	96.46	91.97	90.17	91.06	41.28	40.55	38.73	39.62
LADAN	96.21	91.20	91.51	91.35	56.34	47.76	40.48	42.02
NeurJudge	94.21	88.43	84.73	86.54	37.12	38.16	34.07	36.00
MLK-LJP	97.63	95.84	93.91	95.04	48.15	48.61	46.93	47.83

Accuracy, 0.08 % in Precision, 0.04 % in Recall, and 0.12 % in F1-score over the previous best. On CAIL-small, the improvements are 0.11 % in Accuracy, a substantial 1.49 % in Precision, 0.49 % in Recall, and 0.94 % in F1-score. While these results demonstrate clear performance improvements, the gains for charge prediction are less pronounced compared to the other sub-tasks. This observation suggests that predicting the correct charge relies heavily on the core semantic understanding of the fact description itself, a capability already present to some extent in strong baseline text classification models. Nonetheless, the consistent positive impact across metrics indicates that the fused external legal knowledge, even if providing a more subtle signal for this specific task, does contribute valuable supplementary information, refining the model's ability to pinpoint the appropriate charge.

Finally, we analyze the results for Term of Penalty Prediction, widely regarded as the most challenging LJP sub-task due to its sensitivity to nuanced factual details and quantitative factors. MLK-LJP achieves substantial improvements and establishes new state-of-the-art performance on both datasets for this task. On CAIL-big, it surpasses prior bests by a significant 2.54 % in Accuracy, 0.75 % in Precision, 0.58 % in Recall, and 1.06 % in F1-score. On CAIL-small, the gains are also considerable: 1.03 % in Accuracy, 1.19 % in Precision, 1.85 % in Recall, and 1.39 % in F1-score. We attribute this significant leap in performance primarily to the MLKE module's dedicated capability for extracting multi-modal evidence (specifically, quantitative and image-based knowledge). This is crucial because sentencing decisions often hinge on specific quantitative factors (e.g., monetary values involved in theft or fraud) or physical evidence characteristics (e.g., type of weapon used, visual representation of injuries) - information often present or best represented in non-textual formats. By explicitly extracting and fusing this evidence-related knowledge into the Transformer architecture via our proposed mechanism, MLK-LJP captures critical details that directly influence penalty duration. Notably, none of the other compared methods incorporate a similar targeted mechanism for leveraging multi-modal or quantitative evidence features, highlighting a key advantage and innovation of our approach for this particularly difficult prediction task.

Furthermore, we evaluated the performance of our model on the LAIC2021 dataset, a benchmark widely recognized in the field. The evaluation encompassed two critical judicial tasks: charge prediction and

term of penalty prediction. As shown in Table 6, our proposed MLK-LJP model demonstrates a definitive and significant advantage over all baselines across both sub-tasks. In the charge prediction task, MLK-LJP achieves an F1-score that surpasses the next-best performing model by a substantial margin of 3.69 points. This lead is even more pronounced in the term of penalty prediction task, where our model outperforms the runner-up by 5.81 F1-score points. These results robustly validate the superiority of our framework and its effectiveness in handling complex legal prediction tasks.

5. Analysis

5.1. Ablation analysis

In order to demonstrate the effect of different components in MLK-LJP, we conducted ablation experiments on CAIL-small and CAIL-big and reported the accuracy and F1 values, respectively. The experimental results are shown in Tables 7 and 8. First we need to validate the most important module, the different kinds of legal knowledge acquisition and fusion module. Specifically, in the Tables 7 and 8, “w/o law articles knowledge” refers to removing knowledge of law articles, “w/o legal events knowledge” refers to removing knowledge of legal events, and “w/o quantitative evidence knowledge” refers to the removal of quantitative evidence, “w/o image evidence knowledge” refers to the removal of image evidence, “w/o Legal relation knowledge” refers to the removal of Legal relation Knowledge, where there are the more granular “criminal purpose”, “criminal behavior”, and “post-criminal state”. “w/o All Legal Knowledge” refers to the removal of legal knowledge extraction and multi-modal legal knowledge fusion.” w/o MoLKE” refers to our use of MLP to assign legal knowledge to the transformers.

The experimental results reveal several key insights:

Overall Impact of External Knowledge: Removing the entire legal knowledge module (“w/o All Legal Knowledge”) results in the most substantial performance degradation across all LJP sub-tasks and both datasets. On CAIL-small, Acc/F1 scores dropped by (6.56%/4.9%) for article prediction, (7.98%/8.42%) for charge prediction, and (8.89%/8.94%) for penalty prediction. Similarly significant drops were observed on CAIL-big: (4.26%/4.8%) for articles,

Table 7
Ablation experimental results on CAIL-small.

	Law articles		Charges		Term of penalty	
	Acc.	F1	Acc.	F1	Acc.	F1
MLK-LJP	86.98	86.13	90.14	89.94	45.01	42.82
w/o law articles knowledge	85.77	85.19	89.72	89.04	44.61	41.34
w/o Legal Event Knowledge	85.71	85.24	89.29	88.96	44.69	41.87
w/o Quantitative Evidence Knowledge	85.75	89.94	89.49	89.27	43.72	41.09
w/o Image Evidence Knowledge	85.67	85.29	89.79	89.16	44.53	41.28
w/o Legal Relation Knowledge	85.21	85.19	89.92	89.35	44.19	41.36
-w/o criminal purpose	85.12	85.08	89.71	89.21	44.15	41.31
-w/o criminal act	85.09	85.12	89.67	89.15	44.21	41.07
-w/o post-criminal state	84.92	85.11	89.72	89.31	44.16	41.01
w/o MoLKE	83.25	83.49	87.01	87.62	41.92	39.92
w/o ALL Legal Knowledge	80.42	81.23	82.16	81.52	36.12	33.88

Table 8
Ablation experimental results on CAIL-big.

	Law articles		Charges		Term of penalty	
	Acc.	F1	Acc.	F1	Acc.	F1
MLK-LJP	97.51	88.99	97.92	91.74	68.14	58.12
w/o law articles knowledge	97.19	88.66	97.01	91.09	67.62	57.29
w/o Legal Event Knowledge	97.23	88.81	96.99	91.24	67.71	57.31
w/o Quantitative Evidence Knowledge	97.31	88.29	97.09	91.31	67.18	57.12
w/o Image Evidence Knowledge	97.29	88.33	97.14	91.42	67.15	57.14
w/o Legal Relation Knowledge	97.37	88.67	97.12	91.17	67.79	57.41
-w/o criminal purpose	97.31	88.62	97.09	91.15	67.69	57.39
-w/o criminal act	97.33	88.63	97.12	91.05	67.66	57.38
-w/o post-criminal state	97.33	88.59	97.06	91.11	67.72	57.38
w/o MoLKE	96.17	87.35	96.32	90.94	66.75	56.04
w/o ALL Legal Knowledge	90.25	81.19	90.93	83.24	59.91	49.84

(4.99%/5.5%) for charges, and (5.23%/5.28%) for penalty prediction. This strongly indicates that the externally acquired legal knowledge is highly beneficial and its integration effectively improves overall LJP performance.

Contribution of Individual Knowledge Types: Disabling any single one of the five legal knowledge types consistently leads to a decrease in performance compared to the full MLK-LJP model. This observation validates that each extracted knowledge modality captures distinct and relevant features that contribute positively to the prediction tasks. No single knowledge type appears redundant; rather, they offer complementary information.

Effectiveness of the Knowledge Fusion Mechanism: We compared our sophisticated fusion and projection mechanism against a simpler baseline where a standard Multi-Layer Perceptron (MLP) is used for knowledge integration, replacing core parts of our proposed MoLKE approach. The results clearly show that this simpler MLP approach leads to significant performance degradation across all sub-tasks. On CAIL-small, F1/Acc drops were (2.35%/3.13%) for charges, (2.64%/3.73%) for articles, and (2.9%/3.09%) for penalties. On CAIL-big, the F1/Acc drops were (0.8%/1.6%) for charges, (1.64%/1.34%) for articles, and (2.08%/1.39%) for penalties. This comparison underscores the effectiveness of our specifically designed GNN-based fusion and layer-wise projection mechanism over a more generic MLP integration.

To further probe the benefits of the learning paradigm employed within MLK-LJP, we conducted a comparative experiment on CAIL-small contrasting our prompt-tuning approach against traditional fine-tuning. We trained models using both approaches on training datasets of varying sizes and evaluated their performance consistently on the complete, unseen test set. The comparative results across different data scales for each sub-task are visualized in Figs. 3–5. This comparative analysis yields several key conclusions: (1) **General Superiority of Prompt-Tuning:** Across the entire spectrum of data availability, from extremely limited samples to the full dataset, the prompt-tuning methodology integrated into MLK-LJP consistently outperforms the standard fine-tuning baseline. (2) **Performance on Full Data:** Even when ample training data is available (using the full dataset), MLK-LJP's prompt-tuning maintains a slight performance edge over fine-tuning, as evidenced by the metrics shown in the figures. (3) **Significant Few-Shot Advantage:** Most notably, the advantage of prompt-tuning becomes dramatically apparent in low-data (few-shot) scenarios. This highlights the data efficiency of our chosen approach. Specifically, when trained on only 100 samples, MLK-LJP achieves remarkable F1-score gains compared to fine-tuning: +52.61% for charge prediction, +51.56% for relevant article prediction, and +17.56% for term of penalty prediction.

The collective results from our ablation experiments unequivocally demonstrate the efficacy and contribution of each core component

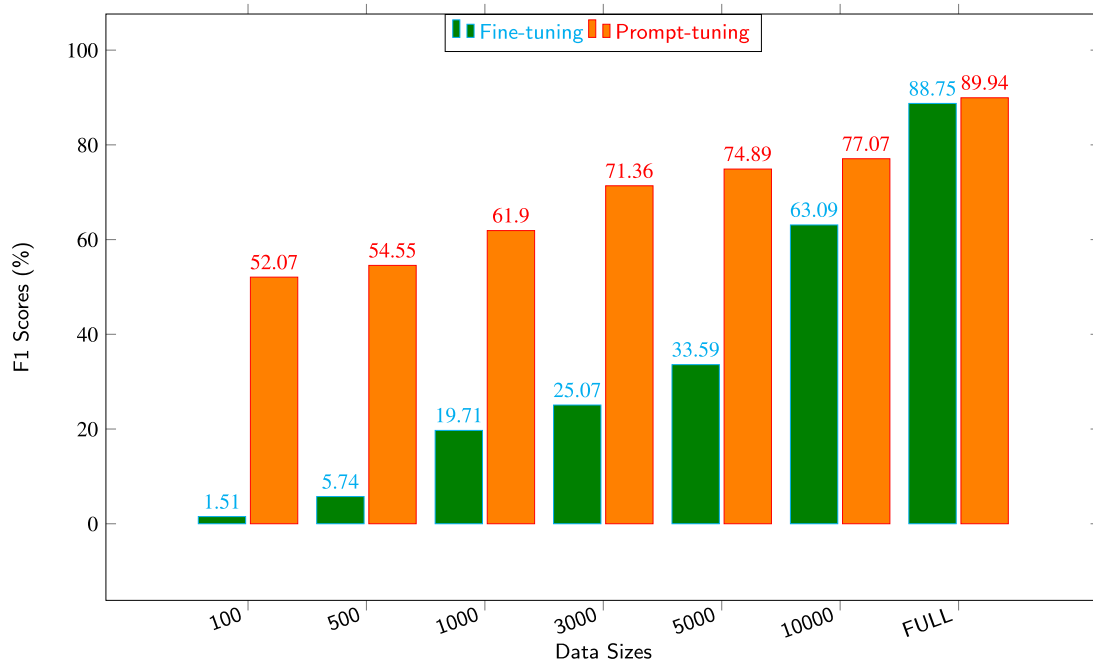


Fig. 3. Charge prediction result of using different data sizes.

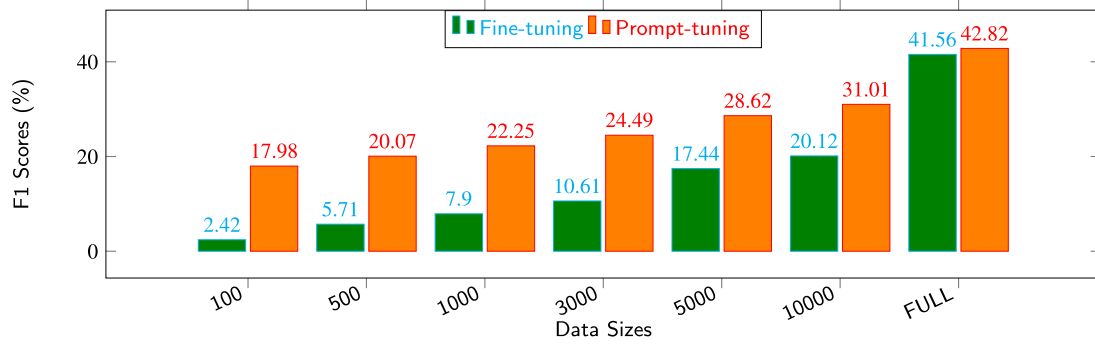


Fig. 4. Term of penalty prediction result of using different data sizes.

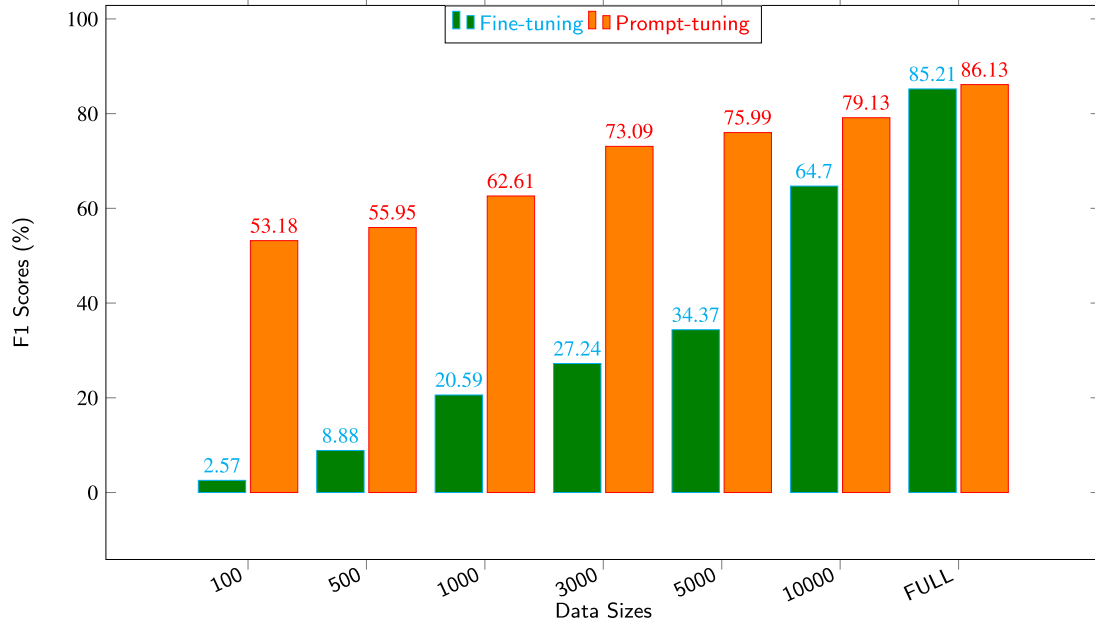


Fig. 5. Law article prediction result of using different data sizes.

within the MLK-LJP framework. The removal of the external knowledge module or any individual knowledge type significantly hampered performance, validating their necessity. Furthermore, the superiority over simpler integration methods and the pronounced advantage of prompt-tuning, especially in few-shot settings, highlight the synergistic power of MLK-LJP's design. It successfully integrates diverse legal knowledge, the representational capabilities of Transformers, and the data-efficient prompt-tuning paradigm, leading to enhanced performance, particularly when labeled data is scarce.

We further conducted comparative experiments on CAIL-small with prompt-tuning and fine-tuning, where Figs. 3–5 show the result of this experiment. We trained the model on different sizes of data from the training set. We evaluate the performance on the same test set shown in Figs. 3–5. We conduct experiments by comparing the performance of fine-tuning and prompt-tuning on different data sizes. The following conclusions can be drawn. (1) Overall, the prompt learning method used by MLK-LJP outperforms the fine-tuning method, both on a small number of datasets and on the entire dataset with sufficient amount of data. (2) On the full dataset, our method performance metrics are slightly ahead of fine-tuning. (3) Our method significantly outperforms fine-tuning on a small number of datasets. In particular, on a dataset of 100 samples, our method achieves 52.61 % higher F1 scores than the fine-tuning method for charges prediction, 51.56 % higher than the fine-tuning method for law articles prediction, and 17.56 % higher than the fine-tuning method for term of penalty prediction.

It can be concluded from the results of the ablation experiments that each component of this paper's approach MLK-LJP is effective and plays an important role. MLK-LJP as a novel knowledge fusion framework organically integrates legal knowledge, Transformer and prompt-tuning. This fusion approach allows the advantages of each module to be taken into account. In addition, the model MLK-LJP achieves performance enhancement under the condition of few-shot dataset.

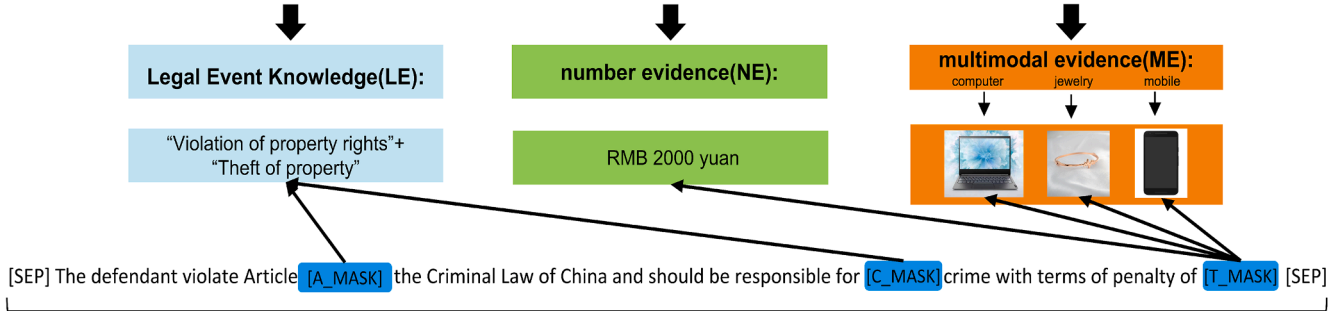
5.2. Interpretability analysis

To gain insights into the decision-making process of MLK-LJP, we conducted interpretability analyses focusing on two complementary aspects: identifying salient proof sentences within the fact description and assessing the contribution of different external legal knowledge components.

Identifying proof sentences via attention. We leverage attention mechanisms inherent in the Transformer architecture to pinpoint sentences in the fact description that most strongly influence the model's predictions. The methodology involves: First, using the trained MLK-LJP model (employing prompt-tuning) to predict the LJP outcomes (charge, article, term) for a given case, where input facts are reconstructed using a prompt template containing [MASK] tokens representing the prediction targets. Then analyzing the multi-head self-attention scores in the final layer of the Transformer encoder. Specifically, for each sentence in the

Fact Description:

Jincheng City District People's Procuratorate alleges that in April 2015, the defendant Song in the urban area of Jincheng City, the theft of parking lots or roadside cars three, the theft of cash 1,000 yuan, computers, jewelry, cell phones and other property, the value of a total of 2,000 yuan. After solving the case, some of the stolen property was recovered and has been returned to the victim.

**Prompt Temple**

	LE	NE	ME1	ME2	ME3
A_MASK	0.69	0.12	0.05	0.17	0.07
C_MASK	0.74	0.27	0.21	0.18	0.05
T_MASK	0.04	0.76	0.52	0.71	0.51



Interpretable output of results	
For charges and articles results:	
Defendant "violation of property rights"+ "theft of property".	
For terms of penalty:	
Relevant quantitative evidence: The property involved is 2,000 yuan.	
Relevant image evidence:	

Fig. 6. First generate multiple knowledge based on the fact description, then do the LJP task using prompt learning, then calculate the attention scores of the corresponding [MASK] positions of charge, law article and term with the five interpretable knowledge, set a threshold to filter the low scores, and finally generate the structured interpretable content.

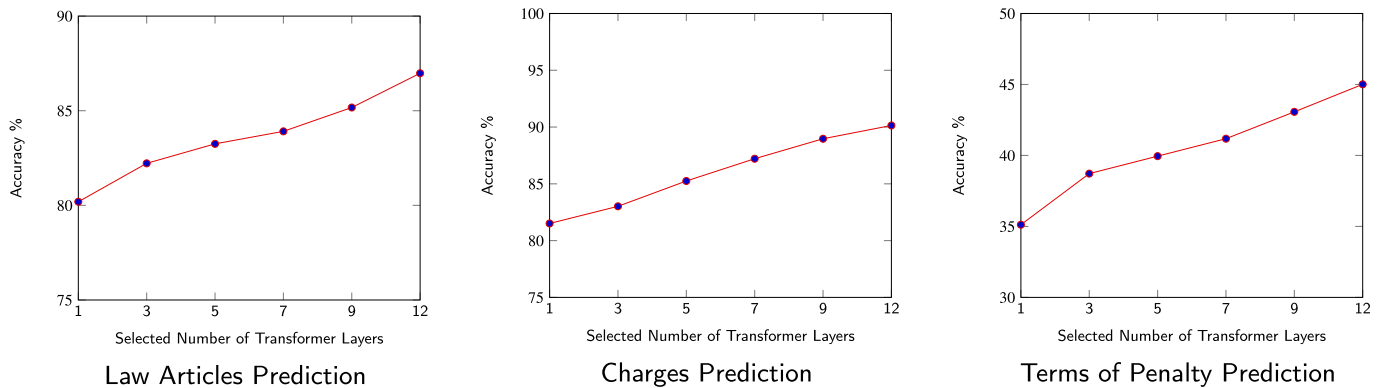


Fig. 7. LJP Results on CAIL-small of MLK-LJP on the effectiveness of using different numbers of layers of text encoder.

fact description, we aggregate the attention scores directed towards the [MASK] token positions across all attention heads. Finally, the resulting aggregated score for each sentence indicates its perceived importance for the overall prediction. These scores are visualized in Tables 9 and 10. For instance, examining Case 1 (Table 9), the analysis reveals that the second sentence ("S2...") receives the highest aggregated attention score. This suggests that the model identified this specific sentence as containing the most critical information for determining the judgment outcome in that case. Similarly, for Case 2, high attention scores on phrases like "with the company subsequently halting production" (in S1) and "failed to deliver the goods" (in S2) align with factual elements strongly indicative of the predicted charge (fraud). This sentence-level analysis helps understand which parts of the factual narrative the model deemed most salient.

Assessing the contribution of external legal knowledge. Beyond sentence-level importance within the facts, we investigate how the integrated external legal knowledge contributes to the predictions. Using a similar attention-based approach, we measure the attention weights assigned by the final prediction representations to the encoded representations of different external knowledge inputs. Due to data characteristics for legal relation knowledge, we focus this analysis on legal event knowledge, quantitative evidence, and image evidence. Specifically, we visualize attention scores from the prediction outputs to these knowledge components, highlighting connections that exceed a predefined threshold (e.g., 0.5, as shown in Fig. 6). Fig. 6 illustrates how the model might, for example, heavily attend to quantitative evidence when predicting the term of penalty, while attending more strongly to legal event knowledge when determining the charge. This knowledge-level interpretability













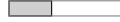





Table 9

We select two cases, and in each case we chose three sentences to demonstrate the attention value.

Cases	Sentences	Content
Case1	S1	...alleged that at 6:00 p.m. on April 22, 2019, the defendant Zhang ...,entered a distribution station while it was unoccupied and stole 2800 RMB from Zhang's wallet...
	S2	Following a police report filed by victim, the defendant, Zhang, was apprehended by police at the distribution station.
	S3	The prosecution contends that the actions of the defendant, Zhang, violated of the Criminal Law of the People's Republic of China, and therefore, he should be charged with the crime of XX.
Case2	S1	It was found during the trial that the sales cooperation agreement concerning Henan Oumou Dairy Biotechnology Co., Ltd. concluded on January 30, 2014, ...with the company subsequently halting production from January 31, 2014 onwards...
	S2	In late November 2014, Chen concealed the cessation of production and, representing Oumou Company, solicited a payment of 24,160 yuan from the victim, Li Mou1, for the purchase of Oumou Company's ...After receiving the payment, Chen failed to deliver the goods to Li Mou1 and, in May 2015, changed his/her mobile phone number, making him/her unreachable by the victim.
	S3	Post-offense, defendant Chen repaid the victim in full, thereby securing the victim's understanding.

Table 10

The Attention Scores (AS) of the sentences in two cases as the evidence of interpretability.

Cases	Sentences	Charges(AS)	Law Articles(AS)	Terms(AS)
Case1	S1	 69.12	 72.38	 64.17
	S2	 0.24	 0.49	 0.75
	S3	 0.37	 0.31	 1.21
Case2	S1	 25.12	 27.98	 35.12
	S2	 37.59	 36.78	 38.67
	S3	 0.58	 0.72	 0.87

provides a complementary perspective, revealing which specific types of external information were most influential for different aspects of the judgment prediction.

In summary, these interpretability analyses, leveraging attention scores from our prompt-based framework, offer valuable insights. They not only identify key evidential sentences within the fact description but also illuminate how MLK-LJP utilizes different facets of the integrated multi-modal legal knowledge to arrive at its predictions, thereby enhancing the transparency and trustworthiness of the model.

5.3. Sensitivity analysis

This section investigates the sensitivity of MLK-LJP's performance to two key hyperparameters: the number of considered image evidences during knowledge extraction and the number of Transformer layers utilized for knowledge fusion.

For number of knowledge fusion layers. Next, we evaluated how performance changes based on the number of Transformer encoder layers into which the fused legal knowledge is integrated. Our default strategy incorporates knowledge across all 12 layers, based on the hypothesis that

this facilitates comprehensive hierarchical integration. To test this, we conducted experiments where knowledge was fused incrementally into the first layer only (Layer 1), then the first two layers (Layers 1–2), up to all twelve layers (Layers 1–12). Figs. 7 and 8 illustrate the results. A clear trend emerges: model accuracy and F1-score generally improve as knowledge is fused into more layers. This validates our hypothesis that integrating knowledge across a sufficient number of layers allows for deeper and more effective interaction between the fact description's semantic representation and the multi-faceted external legal knowledge. Conversely, fusing knowledge into only the initial few layers leads to significantly lower performance, likely because the model has insufficient depth to adequately learn the complex correlations and dependencies between the textual context and the diverse legal knowledge inputs. This underscores the benefit of the deep integration strategy employed in MLK-LJP.

For number of image evidences. we first analyzed the impact of varying the maximum number of image-related evidences extracted and considered by the model. We experimented with values ranging from 1 to 6. The performance trends are presented in Figs. 9 and 10. The results

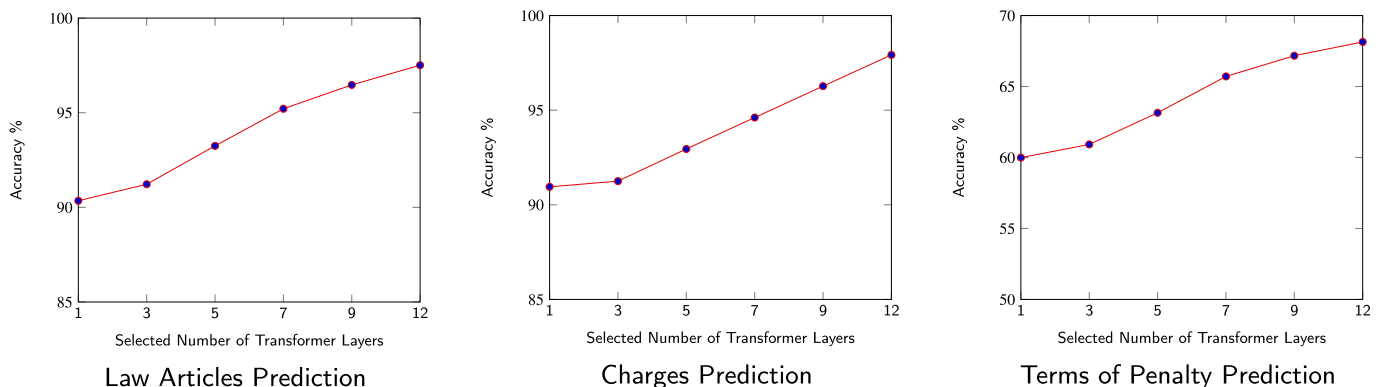


Fig. 8. LJP Results on CAIL-big of MLK-LJP on the Effectiveness of Using Different numbers of Layers of text encoder.

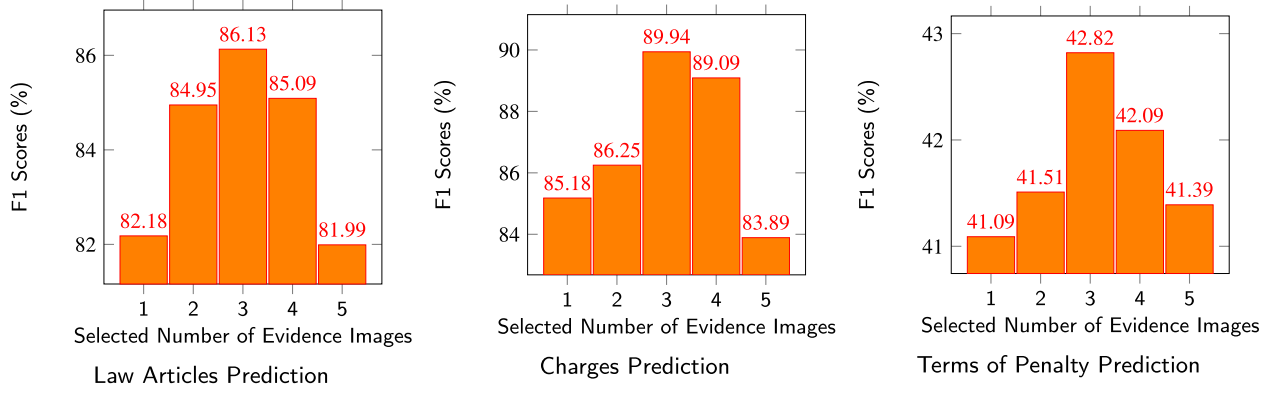


Fig. 9. LJP Results on CAIL-small of MLK-LJP on the Effectiveness of Using Different numbers of evidence images.

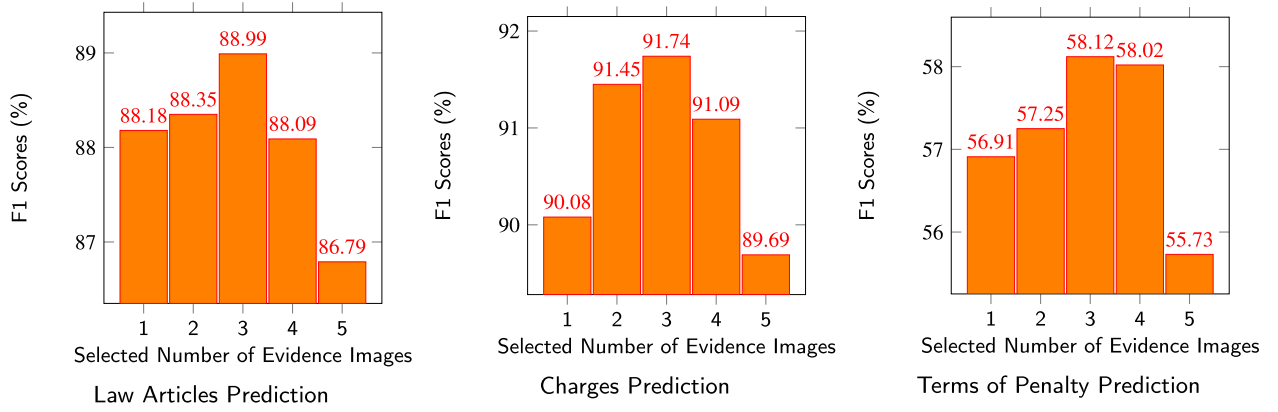


Fig. 10. LJP Results on CAIL-big of MLK-LJP on the Effectiveness of Using Different numbers of evidence images.

Crime of forcible seizing			Crime of picking quarrels and provoking troubles		
Fact Description The People's Procuratorate of Yunxiao County alleges that on September 22, 2014, at approximately 10:00 PM, the defendant, Wu Moujia, drove a motorcycle to the village committee area of Shuangxikou Village in Pumei Town. Taking the victim, Wu Mouyi, by surprise, he snatched the clutch bag tucked under the victim's arm and fled the scene. The bag contained 1,200 RMB in cash, one white Apple iPhone 5, a driver's license, an ID card, and other items. The following day, Wu Moujia sold the stolen phone to another person for 1,800 RMB. An official appraisal determined the value of the said Apple iPhone 5 to be 2,280 RMB. After the incident, the defendant's family compensated the victim, Wu Mouyi, in the amount of 1,500 RMB.			Fact Description The People's Procuratorate of Yuanyang County alleges that on the evening of December 4, 2012, at approximately 11:00 PM, the defendant, Zhang Mougan, armed with a kitchen cleaver, chased and attacked a taxi on a north-south road in Li Moutang Zhuang, Guanchang Township. He then rushed into a restaurant south of the village owned by Sun Moujia and hacked at its door, before proceeding to a gas station where he wounded the owner, Zhou Moumou, in the head and face.		
MPBFN:	crime of robbery	✗	MPBFN:	Crime of willful and malicious injury	✗
LADAN:	crime of robbery	✗	LADAN:	Crime of willful and malicious injury	✗
NeurJudge:	crime of robbery	✗	NeurJudge:	Crime of willful and malicious injury	✗
MLK-LJP(Ours):	crime of robbery	✗	MLK-LJP(Ours):	Crime of willful and malicious injury	✗

Fig. 11. Two cases of error on confusing charges.

indicate that optimal performance is consistently achieved when considering the top 3 image evidences. This observation correlates with the characteristics of the datasets used, where case descriptions frequently mention approximately three distinct items or pieces of evidence potentially representable visually (e.g., weapon type, stolen goods, injury location). Selecting fewer potentially limits relevant information, while selecting more might introduce noise from less relevant items.

5.4. Error cases analysis

We conducted an error case analysis on two cases involving easily confused charges and compared our proposed method against three baseline models. As depicted in Fig. 11, all models rendered incorrect charge predictions for these cases.

In the case on the left, our model, while capable of extracting salient legal concepts, fails to distinguish fine-grained legal nuances, such as the critical difference between force applied to an object versus force applied to a person. It perceives a series of actions—riding a motorcycle, approaching suddenly, snatching, and fleeing—and correlates this entire feature set with a general concept of “forceful theft.” Consequently, it defaults to Robbery, which is the most prominent and severe form of this crime. From a high-level pattern recognition perspective, both Robbery and Snatching share similar features (a perpetrator, a victim, the taking of property, and a swift escape), which leads the model to an erroneous classification.

In the case on the right, the narrative’s central point is the phrase “wounded the owner.” All deep learning-based baseline models place excessive weight on high-impact keywords such as “cleaver,” “chased and attacked,” and “wounded.” Due to this strong keyword correlation, the models infer an intent to harm and misclassify the crime as Intentional Injury. Although our method leverages legal knowledge for assistance, its knowledge granularity is insufficient to comprehend a key legal principle: when an injury occurs within the context of random public disturbance, the primary offense is against public order, not merely the individual. This analytical failure is further compounded by the visual evidence of the kitchen cleaver, which biases the model toward an incorrect prediction.

6. Conclusion

This paper tackled the critical limitation of existing Legal Judgment Prediction methods relying solely on fact descriptions by introducing MLK-LJP, a novel framework designed to incorporate crucial external legal knowledge. MLK-LJP pioneers the integration of multi-granularity, multi-modal knowledge, featuring dedicated modules for extracting five distinct knowledge types (including image evidence) and fusing them effectively using a Graph Neural Network-based MoE approach. To our knowledge, this is the first framework to systematically leverage such diverse multi-modal inputs for the LJP task. Extensive experiments on the CAIL-small and CAIL-big benchmarks confirmed the substantial superiority of MLK-LJP, achieving new state-of-the-art performance across nearly all metrics. Our findings strongly validate the significant benefits of incorporating multi-modal external knowledge, marking a key advancement towards more robust and knowledgeable AI systems in the legal domain.

CRedit authorship contribution statement

Qihui Zhao: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Tianhan Gao:** Writing – review & editing, Supervision, Project administration, Conceptualization; **Nan Guo:** Writing – review & editing, Methodology, Funding acquisition.

Data availability

Data will be made available on request.

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.

References

- [1] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, J. Xu, CAIL2018: a large-scale legal dataset for judgment prediction, CoRR abs/1807.02478 arXiv:1807.02478
- [2] F. Kort, Predicting supreme court decisions mathematically: a quantitative analysis of the right to counsel cases, *Am. Polit. Sci. Rev.* 51 (1) (1957) 1–12.
- [3] W. Lin, T. Kuo, T. Chang, C. Yen, C. Chen, S. Lin, Exploiting machine learning models for Chinese legal documents labeling, case classification, and sentencing prediction, *Int. J. Comput. Linguist. Chin. Lang. Process.* 17 (4) (2012) 49–68. <http://www.aclclp.org.tw/clclp/v17n4/v17n4a4.pdf>.
- [4] Katz, et al., A general approach for predicting the behavior of the supreme court of the united states, *PLoS One* 12 (4) (2014) e0174698.
- [5] Y. Zhang, B.C. Wallace, A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification, in: G. Kondrak, T. Watanabe (Eds.), Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27–December 1, 2017 - Volume 1: Long Papers, Asian Federation of Natural Language Processing, 2017, pp. 253–263. <https://aclanthology.org/I17-1026/>.
- [6] P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning, in: S. Kambhampati (Ed.), Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016, IJCAI/AAAI Press, 2016, pp. 2873–2879. <http://www.ijcai.org/Abstract/16/408>.
- [7] Yao, et al., Graph convolutional networks for text classification, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27–February 1, 2019, AAAI Press, 2019, pp. 7370–7377. <https://doi.org/10.1609/aaai.v33i01.33017370>.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [9] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, M. Sun, How does NLP benefit legal system: a summary of legal artificial intelligence, in: D. Jurafsky, J. Chai, N. Schluter, J.R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, Association for Computational Linguistics, 2020, pp. 5218–5230. <https://doi.org/10.18653/v1/2020.acl-main.466>.
- [10] Wang, et al., Using case facts to predict accusation based on deep learning, in: 19th IEEE International Conference on Software Quality, Reliability and Security Companion, QRS Companion 2019, Sofia, Bulgaria, July 22–26, 2019, IEEE, 2019, pp. 133–137. <https://doi.org/10.1109/QRS-C.2019.00038>.
- [11] Liu, et al., Legal cause prediction with inner descriptions and outer hierarchies, in: M. Sun, X. Huang, H. Ji, Z. Liu, Y. Liu (Eds.), Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings, 11856 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 573–586. https://doi.org/10.1007/978-3-030-32381-3_46.
- [12] Z. Yang, P. Wang, L. Zhang, L. Shou, W. Xu, A recurrent attention network for judgment prediction, in: I.V. Tetko, V. Kurková, P. Karpov, F.J. Theis (Eds.), Artificial Neural Networks and Machine Learning - ICANN 2019: Text and Time Series - 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV, 11730 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 253–266. https://doi.org/10.1007/978-3-030-30490-4_21.
- [13] L. Ma, Y. Zhang, T. Wang, X. Liu, W. Ye, C. Sun, S. Zhang, Legal judgment prediction with multi-stage case representation learning in the real court setting, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11–15, 2021, ACM, 2021, pp. 993–1002. <https://doi.org/10.1145/3404835.3462945>.
- [14] Chen, et al., Charge-based prison term prediction with deep gating network, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, Association for Computational Linguistics, 2019, pp. 6361–6366. <https://doi.org/10.18653/v1/D19-1667>.
- [15] Pan, et al., Charge prediction for multi-defendant cases with multi-scale attention, in: Y. Sun, T. Lu, Z. Yu, H. Fan, L. Gao (Eds.), Computer Supported Cooperative Work and Social Computing - 14th CCF Conference, ChineseCSCW 2019, Kunming,

- China, August 16–18, 2019, Revised Selected Papers, 1042 of *Communications in Computer and Information Science*, Springer, 2019, pp. 766–777. https://doi.org/10.1007/978-981-15-1377-0_59
- [16] Q. Zhao, T. Gao, N. Guo, LA-MGFM: a legal judgment prediction method via sememe-enhanced graph neural networks and multi-graph fusion mechanism, *Inf. Process. Manag.* 60 (5) (2023) 103455. <https://doi.org/10.1016/J.IPM.2023.103455>
- [17] Y. Le, S. Xiao, Z. Xiao, K. Li, Topology-aware multi-task learning framework for civil case judgment prediction, *Expert Syst. Appl.* 238 (Part F) (2024) 122103. <https://doi.org/10.1016/J.ESWA.2023.122103>
- [18] Y. Le, Z. Quan, J. Wang, D. Cao, K. Li, r^2 : a novel recall & ranking framework for legal judgment prediction, *IEEE ACM Trans. Audio Speech Lang. Process.* 32 (2024) 1609–1622. <https://doi.org/10.1109/TASLP.2024.3365389>
- [19] J. Wang, Y. Le, D. Cao, S. Lu, Z. Quan, M. Wang, Graph reasoning with supervised contrastive learning for legal judgment prediction, *IEEE Trans. Neural Netw. Learn. Syst.* (2023) 1–15. <https://doi.org/10.1109/TNNLS.2023.3344634>
- [20] Kang, et al., Creating auxiliary representations from charge definitions for criminal charge prediction, *CoRR abs/1911.05202* (2019). [arXiv:1911.05202](https://arxiv.org/abs/1911.05202)
- [21] Luo, et al., Learning to predict charges for criminal cases with legal basis, in: M. Palmer, R. Hwa, S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017*, Association for Computational Linguistics, 2017, pp. 2727–2736. <https://doi.org/10.18653/v1/d17-1289>
- [22] Hu, et al., Few-shot charge prediction with discriminative legal attributes, in: E.M. Bender, L. Derczynski, P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20–26, 2018*, Association for Computational Linguistics, 2018, pp. 487–498. <https://aclanthology.org/C18-1041/>
- [23] L. Yue, Q. Liu, B. Jin, H. Wu, K. Zhang, Y. An, M. Cheng, B. Yin, D. Wu, Neur-judge: a circumstance-aware neural framework for legal judgment prediction, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event, Canada, July 11–15, 2021, ACM, 2021, pp. 973–982. <https://doi.org/10.1145/3404835.3462826>
- [24] J. Sun, S. Huang, C. Wei, Chinese legal judgment prediction via knowledgeable prompt learning, *Expert Syst. Appl.* 238 (Part E) (2024) 122177. <https://doi.org/10.1016/J.ESWA.2023.122177>
- [25] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing, *CoRR abs/2107.13586* (2021). [arXiv:2107.13586](https://arxiv.org/abs/2107.13586)
- [26] Y. Gu, X. Han, Z. Liu, M. Huang, PPT: pre-trained prompt tuning for few-shot learning, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22–27, 2022, Association for Computational Linguistics, 2022, pp. 8410–8423. <https://doi.org/10.18653/v1/2022.acl-long.576>
- [27] X. Liu, K. Ji, Y. Fu, Z. Du, Z. Yang, J. Tang, P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks, *CoRR abs/2110.07602* (2021). [arXiv:2110.07602](https://arxiv.org/abs/2110.07602)
- [28] J. Wang, C. Wang, F. Luo, C. Tan, M. Qiu, F. Yang, Q. Shi, S. Huang, M. Gao, Towards unified prompt tuning for few-shot text classification, *CoRR abs/2205.05313* (2022). [arXiv:2205.05313](https://arxiv.org/abs/2205.05313) <https://doi.org/10.48550/arXiv.2205.05313>
- [29] S. Hu, N. Ding, H. Wang, Z. Liu, J. Wang, J. Li, W. Wu, M. Sun, Knowledgeable prompt-tuning: incorporating knowledge into prompt verbalizer for text classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22–27, 2022, Association for Computational Linguistics, 2022, pp. 2225–2240. <https://doi.org/10.18653/v1/2022.acl-long.158>
- [30] G. Cui, S. Hu, N. Ding, L. Huang, Z. Liu, Prototypical verbalizer for prompt-based few-shot tuning, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22–27, 2022, Association for Computational Linguistics, 2022, pp. 7014–7024. <https://doi.org/10.18653/v1/2022.acl-long.483>
- [31] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, J. Lu, DenseCLIP: language-guided dense prediction with context-aware prompting, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022*, IEEE, 2022, pp. 18061–18070. <https://doi.org/10.1109/CVPR52688.2022.01755>
- [32] J. Cho, J. Lei, H. Tan, M. Bansal, Unifying vision-and-language tasks via text generation, in: M. Meila, T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research*, PMLR, 2021, pp. 1931–1942. <http://proceedings.mlr.press/v139/cho21a.html>
- [33] L. Cui, Y. Wu, J. Liu, S. Yang, Y. Zhang, Template-based named entity recognition using BART, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1–6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, Association for Computational Linguistics, 2021, pp. 1835–1845. <https://doi.org/10.18653/v1/2021.findings-acl.161>
- [34] D. Lee, A. Kadakia, K. Tan, M. Agarwal, X. Feng, T. Shibuya, R. Mitani, T. Sekiya, J. Pujara, X. Ren, Good examples make a faster learner: simple demonstration-based learning for low-resource NER, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22–27, 2022, Association for Computational Linguistics, 2022, pp. 2687–2700. <https://doi.org/10.18653/v1/2022.acl-long.192>
- [35] Yao, et al., LEVEN: a large-scale Chinese legal event detection dataset, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22–27, 2022*, Association for Computational Linguistics, 2022, pp. 183–201. <https://doi.org/10.18653/v1/2022.findings-acl.17>
- [36] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, H. Chen, Knowprompt: knowledge-aware prompt-tuning with synergistic optimization for relation extraction, in: F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, L. Médini (Eds.), *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25–29, 2022*, ACM, 2022, pp. 2778–2788. <https://doi.org/10.1145/3485447.3511998>
- [37] Xiao, et al., Lawformer: a pre-trained language model for Chinese legal long documents, *CoRR abs/2105.03887* (2021). [arXiv:2105.03887](https://arxiv.org/abs/2105.03887)
- [38] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (6) (2017) 1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
- [39] Xu, et al., Distinguish confusing law articles for legal judgment prediction, in: D. Jurafsky, J. Chai, N. Schluter, J.R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*, Association for Computational Linguistics, 2020, pp. 3086–3095. <https://doi.org/10.18653/v1/2020.acl-main.280>
- [40] Zhong, et al., Legal judgment prediction via topological learning, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31–November 4, 2018*, Association for Computational Linguistics, 2018, pp. 3540–3549. <https://doi.org/10.18653/v1/d18-1390>
- [41] Yang, et al., Legal judgment prediction via multi-perspective bi-feedback network, in: S. Kraus (Ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16, 2019*, ijcai.org, 2019, pp. 4085–4091. <https://doi.org/10.24963/ijcai.2019/567>
- [42] Suykens, Vandewalle, Least squares support vector machine classifiers, *Neural Process. Lett.* 9 (3) (1999) 293–300. <https://doi.org/10.1023/A:1018628609742>
- [43] Kim, Convolutional neural networks for sentence classification, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, ACL, 2014, pp. 1746–1751. <https://doi.org/10.3115/v1/d14-1181>
- [44] Lai, et al., Recurrent convolutional neural networks for text classification, in: B. Bonet, S. Koenig (Eds.), *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25–30, 2015, Austin, Texas, USA, AAAI Press, 2015*, pp. 2267–2273. <http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9745>
- [45] Yang, et al., Hierarchical attention networks for document classification, in: K. Knight, A. Nenkova, O. Rambow (Eds.), *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12–17, 2016*, The Association for Computational Linguistics, 2016, pp. 1480–1489. <https://doi.org/10.18653/v1/n16-1174>
- [46] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, 2019, pp. 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [47] Kingma, Ba, Adam: a method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, 2015. <http://arxiv.org/abs/1412.6980>