Article

Construction of Legal Knowledge Graph Based on Knowledge-Enhanced Large Language Models

Jun Li [1], Lu Qian [2,*], Peifeng Liu [3,*] and Taoxiong Liu [3]

[1] Criminal Justice School, Zhongnan University of Economics and Law, Wuhan 430064, China; lenali@zuel.edu.cn

[2] The School of Transportation and Logistics Engineering, Wuhan University of Technology (WHUT), Wuhan 430063, China

[3] Institute for Innovation and Development, Tsinghua University, Beijing 100084, China; liutx@tsinghua.edu.cn

[*] Correspondence: qianlu@whut.edu.cn (L.Q.); liupeifeng@mail.tsinghua.edu.cn (P.L.)

Abstract: Legal knowledge involves multidimensional heterogeneous knowledge such as legal provisions, judicial interpretations, judicial cases, and defenses, which requires extremely high relevance and accuracy of knowledge. Meanwhile, the construction of a legal knowledge reasoning system also faces challenges in obtaining, processing, and sharing multisource heterogeneous knowledge. The knowledge graph technology, which is a knowledge organization form with triples as the basic unit, is able to efficiently transform multisource heterogeneous information into a knowledge representation form close to human cognition. Taking the automated construction of the Chinese legal knowledge graph (CLKG) as a case scenario, this paper presents a joint knowledge enhancement model (JKEM), where prior knowledge is embedded into a large language model (LLM), and the LLM is fine-tuned through the prefix of the prior knowledge data. Under the condition of freezing most parameters of the LLM, this fine-tuning scheme adds continuous deep prompts as prefix tokens to the input sequences of different layers, which can significantly improve the accuracy of knowledge extraction. The results show that the knowledge extraction accuracy of the JKEM in this paper reaches 90.92%. Based on the superior performance of this model, the CLKG is further constructed, which contains 3480 knowledge triples composed of 9 entities and 2 relationships, providing strong support for an in-depth understanding of the complex relationships in the legal field.

Keywords: knowledge engineering; knowledge extraction; domain knowledge graph; large language model; legal knowledge

1. Introduction

Professional knowledge and skills are highly required for practitioners in the legal field. It generally requires a significant amount of time, manpower, and resources to train qualified and skilled legal practitioners, which is time-consuming and resource-intensive. Legal knowledge inference brings more comprehensive and effective knowledge support tools and decision-making aids for legal practitioners. However, legal knowledge inference involves multidimensional and heterogeneous knowledge such as legal provisions, judicial interpretations, judicial cases, and defense, which has higher requirements for the accuracy of knowledge and the interpretability of reasoning, as well as the ability to analyze the association between knowledge contained in legal provisions and cases. Besides, the construction of a legal knowledge inference system faces increased challenges due to difficulties such as the acquisition, processing, and sharing of multisource heterogeneous knowledge. Therefore, how to construct CLKG and automatically extract accurate and reliable knowledge from multiple legal sources has important theoretical significance and practical value in the legal field.

With the continuous development of artificial intelligence technology in industry and academia, knowledge graph technology keeps drawing attention due to its unique

advantages in the field of vertical knowledge management [1]. Since knowledge graphs [2] are organized based on knowledge triplets, they show natural advantages in effectively organizing and logically structuring knowledge from heterogeneous multiple sources [3], which has been widely applied to common-sense reasoning [4,5] and specific domain reasoning tasks [6,7] with excellent performance. The structure of knowledge triplets can be summarized as (e_1, R, e_2), where e_1 represents the source entity, e_2 denotes the target entity, and R signifies the semantic relationship between e_1 and e_2. Entities embody particular objects, ideas, informational resources, or data pertaining to them, whereas semantic relationships signify the linkages among these entities [8].

Building a knowledge graph involves two primary steps: named entity recognition (NER) [9] and relation classification (RC) [10]. In traditional methods, these two subtasks are relatively independent, regardless of the correlation between the two subtasks, which not only loses a lot of relevant contextual semantic information, but also causes error propagation problems [11]. To address this problem, inspired by the idea of the joint neural network knowledge extraction based on the transformer architecture, this paper proposes an enhanced model for legal domain knowledge to automatically construct the knowledge graph in this field. To this end, prior knowledge is embedded into an LLM, which is fine-tuned with the prefix of prior knowledge data to construct the knowledge-enhanced model. This prefix-tuning approach, which freezes most parameters of the LLM, incorporates consecutive deep prompts as prefix tokens into the input sequences of different layers.

The main contributions of this paper include the following aspects:

(1) Based on the principle of legal ontology consensus reuse [12,13], the structure of the CLKG is constructed combined with the professional knowledge of legal experts, where nine entities and two relationships are defined. The corpus of knowledge in this paper is derived from the Criminal Code of the People's Republic of China and Annotated Code (New fourth edition) [14], supplemented by Internet encyclopedia data.

(2) A JKEM is proposed based on prior knowledge embedded in LLM, which is fine-tuned with the prefix of prior knowledge data. The model demonstrates high recognition performance independent of (on the condition of not depending on) artificially set features, where the corresponding knowledge triples can be correctly extracted from legal annotations in natural language, achieving a knowledge extraction accuracy of 90.92%.

(3) Based on the superior performance of the legal JKEM, a knowledge graph of Chinese legal is constructed, which contains 3480 pairs of knowledge triples composed of 9 entities and 2 relationships. It provides structured knowledge information to further facilitate legal knowledge inference.

2. Related Works

The legal field is a domain that heavily relies on knowledge and experience, thus knowledge management and reasoning in this field have always been the focus of cross-disciplinary research [15]. In recent years, the use of artificial intelligence technology in the judicial field has been widely discussed in academia, including the advantages and disadvantages of artificial intelligence in legal judgments. The entry of artificial intelligence into the legal field can contribute to "fair justice" or discretionary moral judgment [16]. However, there is insufficient attention paid to whether artificial intelligence methods conform to the values, ideals, and challenges of the legal profession [17]. With the emergence of GPT technology, its outstanding performance in both subjective and objective legal fields [18]. Traditional legal document analysis relies on manual reading or traditional statistical machine learning models, and its efficiency and accuracy need to be improved [19]. The legal document analysis technology based on LLMs, with its excellent semantic understanding ability and contextual modeling ability, can effectively solve these problems and provide powerful auxiliary tools for legal practitioners [20]. LLMs can help automate various aspects of legal document analysis, including key information extraction, risk identification, and argument analysis [21].

Therefore, cross-disciplinary research on law and computer technology based on graph-structured knowledge representation and management has drawn increasing attention in recent years. The Lynx project, funded by the European Union [22], focused on creating a CLKG and applying it to the semantic processing, analysis, and enrichment of legal domain documents. Tong et al. [23] used graph-structured data for legal judgment prediction in the civil law system, which is a multitask and multilabel problem involving the prediction of legal provisions, charges, and penalty provisions based on factual descriptions. Bi et al. [24] applied CLKG to legal charge prediction in legal intelligence assistance systems. Automatic legal charge prediction aims to determine the final charge content according to the factual description of criminal cases, which can assist human judges in managing workload and improving efficiency, provide accessible legal guidance for individuals, and support enterprises in litigation financing and compliance monitoring. In light of the above, the construction and application of CLKG have become an important and popular research field.

In the process of constructing a CLKG, the extraction of the entity and relationship is an essential step in constructing the corresponding knowledge base [25]. In order to achieve the automatic construction of the CLKG, it is necessary to establish an efficient model that can automatically extract entities and relationships. Highly advanced deep neural network (DNN) models exhibit exceptional proficiency in handling such tasks, contributing significantly to the establishment of knowledge graphs. These DNN models excel in delivering accurate results for both NER and RC. Nevertheless, a prevalent issue in current technical systems is the treatment of NER and RC as distinct tasks, each requiring a separate model. This approach not only demands substantial training resources but also leads to the accumulation of errors [26]. As LLMs exhibit outstanding performance in the field of knowledge organization, representation, and reasoning, their excellent capabilities can be utilized to automatically extract knowledge entities and relationships jointly from unstructured text, thus achieving end-to-end construction of knowledge graphs in specific professional domain.

It can be found that considerable efforts have been made to construct knowledge graphs with LLMs. Pan et al. [27] summarized and prospected the technical routes for the integration of knowledge graphs and LLMs technology. It can be seen that the end-to-end construction of knowledge graphs with LLMs can be generally summarized into three categories: 1. Utilizing soft prompts and contrastive learning methods, the knowledge graph is constructed and completed through prompts engineering [28]. 2. Adopting a pipelined knowledge graph construction scheme, where the entities of the knowledge graph are generated by the LLM, followed by a relationship construction, so that the knowledge graph can be efficiently constructed from the text description [29]. 3. The fine-tuning LLM is used as an encoder for the representation and completion of knowledge [30]. This paper analyzes and summarizes the strengths and weaknesses of the above three methods, as shown in Table 1 [31–34].

Table 1. Comparison of the technologies for end-to-end constructing knowledge graphs with LLMs.

Methods	Advantages	Disadvantages
Soft Prompts	1. No additional training: Immediate utilization without the need for further model training. 2. Rapid Deployment: Enables swift implementation due to the absence of a training phase.	1. Significant Randomness: Outputs can vary widely, leading to inconsistency. 2. Uncontrollable Results: Lack of precise control over the generated outputs.
Pipelined Scheme	1. High Technical Maturity: Well-established methods with a solid foundation in existing research and practice. 2. Modular Approach: Allows for independent optimization of each stage in the pipeline.	1. Labeled Data Requirement: Necessitates labeled training samples, which can be costly and time-consuming to obtain. 2. Cumulative Errors: Errors from each stage can propagate and accumulate, affecting final output quality.

Methods	Advantages	Disadvantages
Fine-tuning LLMs	1. Strong Consistency: Ensures coherence and uniformity in the generated knowledge graph. 2.Error Mitigation: Avoids cumulative errors by integrating knowledge directly into the model.	1. Labeled Data Requirement: Similarly, requires labeled training data, which may limit applicability in certain domains. 2. Heavy Training: Involves extensive computational resources and time for fine-tuning large models.

3. CLKG Construction and Management Framework

According to the technical routes summarized in Table 1, the scheme of fine-tuning the LLM is adopted in this paper. To be specific, the prior knowledge is embedded into the LLM, and a knowledge-enhanced model is constructed by fine-tuning the LLM with the prefix of the prior knowledge data to construct. Then, the continuous deep prompts, as prefix tokens, are added to the input sequences of different layers while the most parameters of the LLM are frozen.

In order to realize the automatic construction of CLKG, this paper developed the corresponding knowledge acquisition and management software, whose framework is shown in Figure 1.
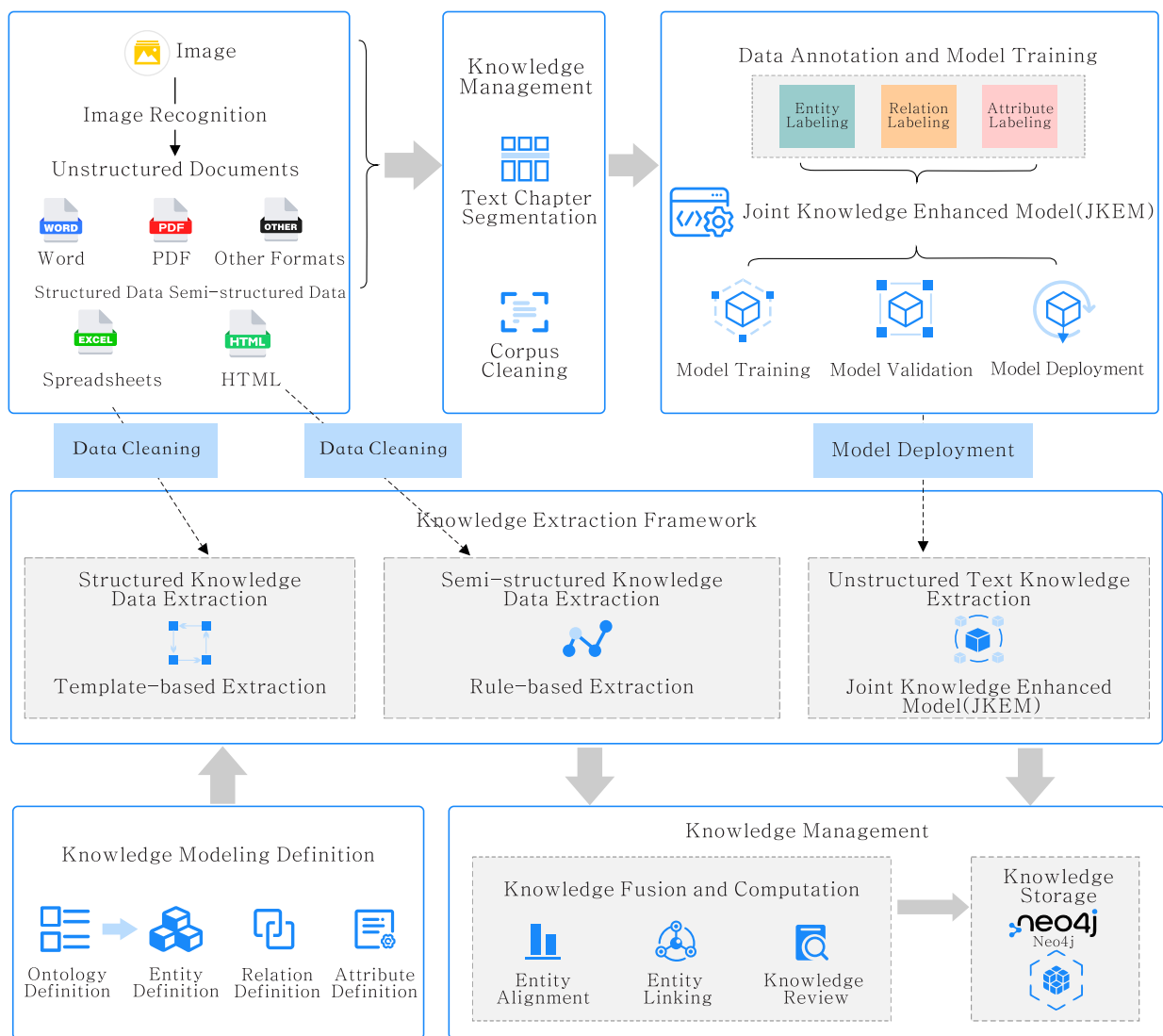

Figure 1. Legal knowledge acquisition and management framework.

From Figure 1, it can be seen that, first, it is necessary to define the knowledge structure in terms of the entities, relationships, and relevant properties required for the CLKG construction. Secondly, the data sources need to be classified, which are divided into unstructured text data, structured table data, and semistructured data in this paper. Different methods are then used to extract knowledge from different types of data. Specifically, structured table data can be extracted through simple template extraction, semistructured data can be extracted through rule-based methods, and unstructured data can be extracted through the knowledge-enhanced model constructed in this paper. Finally, an accurate and reliable CLKG is constructed with the extracted triple data after knowledge alignment, linking, and review.

3.1. The Knowledge Sources of CLKG

The corpus used in this paper comes from "The Criminal Law of the People's Republic of China: Annotated Code (New Fourth Edition)" [14]. It provides comprehensive and systematic interpretations of the law of the People's Republic of China, including related laws and regulations. This book not only includes the original text of the legal but also provides in-depth analysis and explanation of each clause through annotations, interpretations, and case studies. Additionally, in order to enrich the data types, this paper supplements the corpus with data from internet encyclopedias.

3.2. The Framework of CLKG

The universality and reusability of knowledge are definitely considered in the construction of the CLKG framework. However, there is no unified knowledge graph structure standard in the legal field so far. Therefore, this study combines expert knowledge in the legal field and related books, documents, and other materials and reuses the ontology consensus of existing related field knowledge bases [12]. Regarding crime as the core entity, nine entities are defined in this paper, namely, crime entity, concept entity, constitutive characteristic entity, judging standard entity, punishment entity, legal provision entity, judicial interpretation entity, defense entity, and case entity. The corresponding explanation and examples of each entity are shown in Table 2. Since the examples of some entities are too long, only part of the content is shown in the table.

It can be seen from Table 2 that the entities defined in this paper have important reuse value in actual legal activities, and the knowledge information required for legal activities can be completely restored through the definition of nine kinds of entities. Meanwhile, it can be seen that the entity definitions in the CLKG are different from those in the encyclopedic knowledge graph, where the text corresponding to some entities is relatively long in order to retain the integrity of legal knowledge.

Once a distinct entity classification system is in place, the connections among corresponding entities can be accurately delineated. In this paper, we outline two types of relationships, providing brief introductions and examples for each:

(1) Entity–With (EW): Entities with relationships.
Example: The concept of the crime of damaging environmental resources protection, the constitutive features of the crime of damaging environmental resources protection, etc.

(2) Component–Whole (CW): Relationships between the whole and its components.
Example: The crime of damaging environmental resources protection includes the crime of major environmental pollution accidents, the crime of endangering public safety includes the crime of placing dangerous substances and the crime of damaging transportation vehicles, etc.

Building on the aforementioned, this paper introduces a framework comprising nine entity types and two relationship types within the realm of legal knowledge. As illustrated in Figure 2 architecture diagram, these components are utilized to construct triples for domain-specific knowledge graphs. The resultant knowledge graph comprehensively showcases the intricacies and logical connections of legal knowledge.

Table 2. Definition and Examples of Legal Knowledge Graph Entities.

Entity Names	Entity Definitions	Entity Examples
Crime entity	The crime in criminal law	Crimes endangering national security, treason against the state, etc.
Concept entity	The definition of the crime	The crime of treason against the state refers to the act of colluding with foreign countries or overseas institutions, organizations, or individuals to endanger the sovereignty, territorial integrity, and security of the People's Republic of China.
Constitutive characteristic entity	Object elements constituting a crime	The object of this crime is the People's Republic of China, its territorial integrity, and security.
Judging standard entity	The standard for determining the crime and the difference from other crimes	The boundary between this crime and the crime of subverting state power, etc.
Punishment entity	The intensity of punishment	Whoever commits this crime shall be sentenced to life imprisonment or imprisonment of ten years or above. According to the provisions of Articles 56 and 113 of this Law, whoever commits this crime shall be additionally deprived of political rights and may be concurrently sentenced to confiscation of property.
Legal provision entity	Original legal text	Article 102 Whoever colludes with a foreign State to endanger the sovereignty, territorial integrity and security of the People's Republic of China shall be sentenced to life imprisonment or fixed-term imprisonment of not less than 10 years. etc.
Judicial interpretation entity	The interpretations made by the highest judicial organ of the state on specific application legal issues in the process of applying the law.	The Supreme People's Court and the Supreme People's Procuratorate issued the "Interpretation on Several Issues Concerning the Specific Application of Laws in Handling Criminal Cases of Organizing and Utilizing Cult Organizations" (1999, 10, 20 Interpretation No. 19 [1999]), etc.
Defense entity	Sample defense statements based on actual cases	Dear Chief Judge and Judge, entrusted by the relatives of the defendant and appointed by Beijing Jietong Law Firm, I am the defender of the defendant Chen. Based on the facts and cross-examination evidence of the trial investigation and in accordance with the relevant laws of our country, the following defense opinions are expressed, etc.
Case entity	Representative cases that actually occurred	In his statement on 20 July 2009, Xu xx admitted to receiving the wool sweaters from Sifang Company, and witness Li xx (Chairman of Sifang Yitong Automobile Trading Co., Ltd.) also mentioned in his statement on 21 April 2009 that he sent wool sweaters, mobile phones, computers and other items to the two defendants in this case, etc.

Figure 2. The architecture diagram of entities and relationships in the legal knowledge graph.

3.3. Knowledge Annotation Scheme and Tools for CLKG

When in-depth analyzing the entity and relationship structure of the CLKG, the paper clearly points out that as a bridge connecting different entities, the existence of a relationship is strictly limited to between entities, and nonentity elements do not have the ability to carry relationships. Based on this understanding, entities are naturally divided into two categories: one is the entities that participate in building clear relationships, and the other is the isolated entities that do not directly participate in the relationship construction.

Under this theoretical framework, this paper designs a comprehensive annotation scheme that covers the following three key aspects to ensure comprehensive and accurate annotation of the legal knowledge corpus:

(1) Fine annotation of entity location information: The BMEO (Begin, Middle, End, Other) character-level annotation strategy is adopted, where B is used to mark the starting position of the entity, M follows to represent the continuous characters within the entity, E marks the end of the entity, and O is applied to ordinary text characters of non-named entities. In this way, the entity boundaries can be accurately located and distinguished.

(2) Detailed classification of entity category information: Nine specific entity categories are defined in this paper, which comprehensively cover the key elements and concepts in the legal knowledge graph. The specific classification details are shown in Table 2, which provides a rich semantic dimension for in-depth analysis.

(3) Definition of entity relationship information: In order to deeply understand the interactions and connections between entities, this paper further clarifies two core relationship types, which form the basis of the complex associations between entities in the legal knowledge graph.

Subsequently, the above annotation scheme is adopted to systematically annotate the corpus data. In order to enhance the efficiency of annotation and the visual representation of model performance, this study conducted in-depth customization and secondary development based on the mature text annotation platform Doccano [35], aiming to build a more suitable human–machine interaction tool for progressive and continuous annotation tasks. The core innovation of this tool lies in the integration of pretrained models, which enables preliminary automatic annotation of entities and relationships, transforming the traditional manual annotation process into efficient review and verification tasks. This transformation not only greatly improves the efficiency of the annotation work, but also effectively reduces human errors. Furthermore, it enhances the transparency and reliability of the annotation process through an intuitive display of the annotation results. This paper annotated legal knowledge entities related to 460 criminal offenses in the Criminal Code. The statistics of entities and relationships are shown in Table 3.

Table 3. Statistics on the number of entities and relationships in the legal knowledge graph.

Entity or Relationship Category	Number of Entities/Relationships
Crime Name Entities	460
Concept Entities	435
Constituent Feature Entities	433
Identification Standard Entities	394
Punishment Entities	412
Legal Provision Entities	417
Judicial Interpretation Entities	237
Defense Statement Entities	169
Case Entities	169
Total Number of Entities	2957
Entities with Relationships	435
Whole–Part Relationships	3045
Total Number of Relationships	3480

4. The Framework of JKEM

4.1. Overview of JKEM

The JKEM is based on prefix fine-tuning of the LLM [36], which is a fine-tuning approach for pretrained models. It introduces a trainable "prefix" sequence, namely a series of learnable embedding vectors, before the input layer of the pretrained model to influence the model outputs without modifying the parameters of the pretrained model itself. Inspired by the prompting method, but prefix-tuning differs by using continuous, trainable vectors as prefixes, which are optimized during training to guide the language model to generate correct outputs. This approach adapts the pretrained language model to downstream tasks by adjusting a few task-specific parameters while keeping the pretrained language model parameters fixed. The difference from traditional prefix-tuning is that this study embeds the prefix into each layer of the LLM and constructs a knowledge enhancement model by prefix fine-tuning the LLM, as shown in Figure 3.
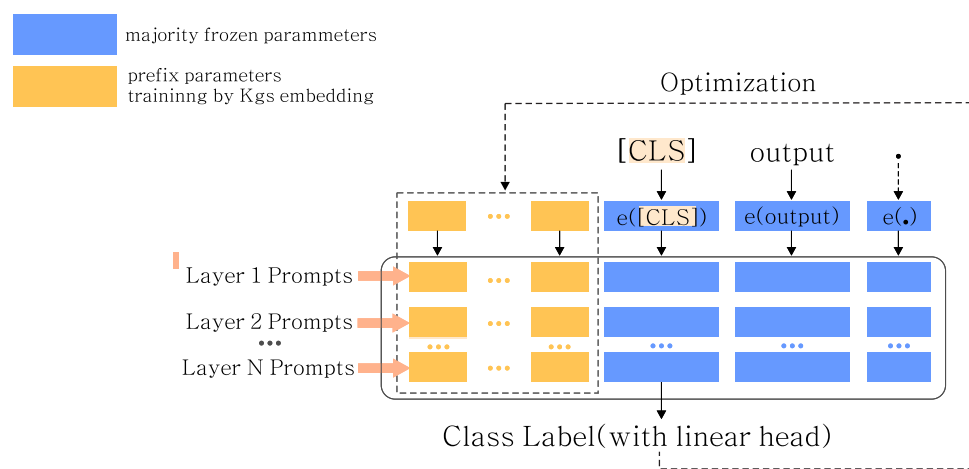

Figure 3. The architecture of JKEM.

In Figure 3, the blue parts are the frozen large model parameters θ, while the deep yellow part is the prefix parameters p that need to be trained. The number of fine-tuning parameters p does not exceed 0.3% of the large model parameters θ. In this way, the original intention of extracting preliminary legal knowledge triples from professional texts is achieved by embedding the LLM with prior knowledge and fine-tuning a small number of parameters.

4.2. Tasks Description

The input of a pretrained language model is a sequence $x = [x_1, x_2, \cdots, x_n]$, where x_i is the embedding representation of the *i*-th element in the sequence (e.g., word embedding). In prefix-tuning, we add a prefix sequence $p = [p_1, p_2, \cdots, p_m]$ to the input sequence, where $p_j \in \mathbb{R}^d$ is the *j*-th learnable embedding vector in the prefix, *d* is the dimension of the embedding vector. Therefore, the input that the model actually receives is the concatenation of the prefix and the original input sequence following Equation (1).

$$[p, x] = [p_1, p_2, \cdots, p_m, x_1, x_2, \cdots, x_n] \tag{1}$$

In the Transformer model, the input passes through multiple self-attention layers. In each self-attention layer, the input is linearly transformed into query (Q), key (K), and value (V) matrices. However, in prefix-tuning, we only train the prefix embeddings and keep the rest of the model unchanged, such as the linear transformation weights of the query, key, and value matrices.

This new input sequence as Equation (1) is then fed into the pretrained language model, which performs a series of processing, including a self-attention mechanism and

feedforward network, to generate the output sequence y. The calculation process is abstracted into a function f, representing the mapping of the model from input to output. Therefore, the process of prefix-tuning can be expressed as Equation (2),

$$y = f([p, x], \theta), \tag{2}$$

where θ represents all parameters in the pretrained language model except prefix embedding, which are fixed during the training process. The prefix embedding p is trainable, and it will be optimized during the training process to minimize a certain loss function L. In this paper, the input x is professional legal text knowledge, and the output y is the extracted knowledge triple.

4.3. Objective Function

To optimize the prefix embeddings for a specific task, a loss function L is defined, which measures the difference between the model output y and the true labels t. The training process aims to find the optimal prefix embeddings p^* that minimize this loss function L as Equation (3).

$$p^* = \arg \min_p L(f([p; x], \theta), t), \tag{3}$$

where f in the formula is a highly abstract function that actually represents a series of complex operations in the pretrained language model, including multiple self-attention layers, feedforward network layers, layer normalization, etc. In addition, the specific form of the loss function L will also vary depending on the task. In this paper, the task is text generation tasks, thus negative log-likelihood loss is used. During training, gradient descent optimization algorithms are applied to update the prefix embedding p while keeping θ unchanged. By iteratively adjusting p, we can gradually adapt the model to specific downstream tasks. In this paper, the specific downstream task is the extraction of knowledge triplets.

5. The Validation of JKEM

5.1. Experiment Setting

The framework development environment in this study is based on Windows 10, where the system type is a 64-bit operator, the CPU is Intel Core i9-11950H, the memory is 32 G, and the GPU is RT XA4000 8 G. Python 3.8.8 is used for development. The graph database uses Neo4j with the version 4.3.4.

In the knowledge enhancement model, the prior knowledge triples are embedded in the LLM, where ChatGLM-6B [37] is adopted. Since the input and output of the model are Chinese texts, more tokens are needed to complete the tasks. This paper sets the length of natural language instructions PRE_SEQ_LEN to 256, the maximum length of the input sequence max_source_length to 512, and the maximum length of the output sequence max_target_length to 512.

5.2. Corpus Information

In order to systematically construct a CLKG with professional depth, this study uses "The Criminal Law of the People's Republic of China: Annotated Code (New Fourth Edition)" [14] as the core knowledge source, supplemented by internet data as auxiliary information sources. A legal knowledge corpus has been finely created, which aims to comprehensively cover the core concepts and logical relationships in the legal field, laying a solid foundation for subsequent knowledge extraction and graph construction.

In the specific implementation process, in order to ensure the scientificity and rationality of model training, verification, and evaluation, we divided the legal knowledge corpus into a training set, a verification set, and a test set according to the standard ratio of 10:1:1, aiming to balance the demands of the model's learning ability and generalization ability

evaluation. More importantly, while maintaining the randomness of data distribution, we pay special attention to the coexistence of 9 core entity types and 2 key relationship types in the verification set and the test set. This is to effectively improve the generalization performance of the model by enhancing the diversity of the dataset and to ensure the comprehensiveness and effectiveness of the test results, so as to more accurately reflect the application ability of the model in unknown or complex legal scenarios. This is to effectively improve the generalization performance of the model by enhancing the diversity of the dataset and to ensure the comprehensiveness and effectiveness of the test results, so as to more accurately reflect the application ability of the model in unknown or complex legal scenarios. Furthermore, for the test set, detailed statistical analyses of the number of entities and relationships are conducted, which are summarized in Table 4.

Table 4. Statistics on the number of entities and relationships in the corpus test set.

Entity or Relationship Category	Number of Entities/Relationships
Crime Name Entities	35
Concept Entities	35
Constituent Feature Entities	35
Identification Standard Entities	34
Punishment Entities	35
Legal Provision Entities	34
Judicial Interpretation Entities	19
Defense Statement Entities	15
Case Entities	15
Entities with Relationships	35
Whole–Part Relationships	254

The statistical results in Table 4 not only intuitively demonstrate the composition characteristics of the test set data, but also provide a quantitative basis for our in-depth analysis of the model performance, which ensures a comprehensive and objective evaluation of constructing the legal knowledge graph.

5.3. Experimental Results and Comparison of the JKEM on the Legal Knowledge Corpus

The JKEM proposed in this paper was compared with other classic models on the legal knowledge corpus, including the conditional random field (CRF) model of statistical learning [38], the deep learning-based bidirectional long-short term memory (BiLSTM) model [39], the BERT [40] model based on the Transformer [41] architecture, and the untuned LLM ChatGLM-6*B* [37]. The results of the comparative experiment are shown in Table 5.

Table 5. The comparative results on the legal knowledge corpus.

Types of Model	Accuracy/%	Recall/%	F1 Value/%
CRF	78.65	75.60	77.09
BiLSTM	82.30	83.16	82.73
BERT	85.72	84.91	85.31
ChatGLM-6B	86.20	85.92	86.06
JKEM(Ours)	90.78	91.06	90.92

From the experimental results shown in Table 5, it is evident that the JKEM model is significantly superior to other comparison models in terms of accuracy, recall, and F1 value. Figure 4 is drawn according to Table 5 for the intuitive display.
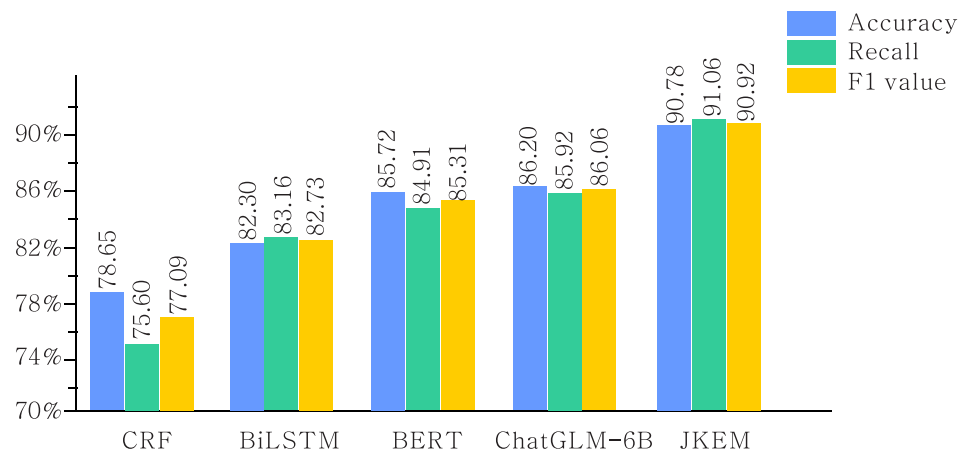
Figure 4. Model performance comparison.

As a representative of traditional statistical learning methods, the CRF model, despite having some application value in specific tasks, shows relatively limited performance when dealing with complex legal texts (accuracy 78.65%, recall 75.60%, F1 value 77.09%). In contrast, the BiLSTM model can capture long-distance dependencies in the text by introducing deep learning technology, thus achieving a greatly performance improvement (accuracy 82.31%, recall 83.17%, F1 value 82.74%). With the rise of pretrained language models, the BERT model performs even better on the legal knowledge corpus with its powerful language representation ability and extensive contextual understanding ability (accuracy 85.74%, recall 84.91%, F1 value 85.32%). As an example of an LLM, the ChatGLM-6B model, although not fine-tuned in a specific field, still performs well in legal text processing due to its powerful language generation and understanding ability (accuracy 86.23%, recall 85.92%, F1 value 86.07%), further demonstrating the potential of large-scale pretrained models in cross-domain applications. The JKEM model has achieved the best performance in all indicators. By integrating external legal knowledge bases and domain-specific knowledge, the model significantly enhances the understanding and reasoning ability in the legal field, thus achieving excellent performance of 90.76% accuracy, 91.05% recall, and 90.90% F1 value. This result not only verifies the effectiveness of knowledge enhancement strategy in improving model performance, but also provides new ideas and methods for research in the field of legal knowledge processing.

In addition, we thoroughly explore the predictive capabilities of our proposed JKEM model on various legal entities and relationships in the legal knowledge corpus. Through systematic experimental design and analysis, the detailed experimental results of the JKEM in the test set are obtained as shown in Table 6, which summarizes the model's precision, recall, and F1 value when identifying different legal entities and relationships, providing a comprehensive perspective for evaluating the model's performance.

Table 6. Prediction results of various entities and relations on the legal knowledge corpus.

Category	Accuracy/%	Recall/%	F1 Value/%
Crime	100.00	97.14	98.55
Concept	91.43	94.29	92.84
Characteristic	88.57	91.43	89.98
Judging standard	97.06	91.18	94.03
Punishment	91.43	94.29	92.84
Legal provision	97.06	88.24	92.44
Interpretation	89.47	89.47	89.47
Defense	100.00	100.00	100.00
Case	80.00	86.67	83.20
Entity–With	91.43	94.29	92.84
Component–Whole	98.82	97.24	98.03

The entity prediction results are analyzed as follows:

(1) Crime entity: The JKEM model has achieved excellent results in the recognition of crime entities, with an accuracy and recall of 100.00% and 97.14%, respectively, and an F1 value as high as 98.55%. which fully demonstrates the model's strong ability to accurately capture the core crime information in legal texts.

(2) Concept entity, punishment entity, and legal provision entity: For these three key legal entities, the model also shows good prediction performance, with the accuracy and recall rates remaining at a relatively high level, and the F1 scores stably ranging from 92% to 94%. This indicates that the JKEM model has significant advantages in understanding and distinguishing legal concepts, punishment measures, and legal provisions.

(3) Constitutive characteristic entity and judging standard entity: Although the prediction of these two types of entities is relatively more challenging, the model still achieved relatively satisfactory results, with F1 values reaching 89.98% and 94.03%, respectively, which reflects the robustness of the model in dealing with complex features and identification standards in legal texts.

(4) Judicial interpretation entity and defense entity: The model also performed well in the recognition of judicial interpretation and defense entities, especially the recognition of defense entities achieved perfect accuracy and recall (both 100%). This further verifies the ability of the JKEM model to capture highly specialized and precision-demanding content in legal texts.

(5) Case entity: Compared with other entity types, the recognition effect of case entity is slightly insufficient, with accuracy and recall rates of 80.00% and 86.67%, respectively, and an F1 value of 83.20%. This may be related to the diversity and complexity of case descriptions, suggesting that the model needs to be further optimized in the future to better handle such texts.

With regards to relationship prediction results:

(1) Entity–With relationship: When identifying the basic association relationship between entities, the model also performed well, with high accuracy and recall (91.43% and 94.29%, respectively), and F1 value of 92.84%, which shows that the JKEM model has good ability in understanding the basic relationship between entities in legal texts.

(2) Component–Whole relationship: In the more sophisticated semantic relationship recognition task, i.e., the recognition of the whole–part relationship, the model demonstrated extremely high performance, with accuracy and recall reaching 98.82% and 97.24%, respectively, and F1 value as high as 98.03%. This result fully demonstrates the superiority of the model in capturing the complex semantic structure in legal texts.

In order to verify the fitting of the model and avoid overfitting or underfitting, this paper calculated the loss of 30 epochs, as shown in Figure 5.

Figure 5 shows in detail the changes in model training loss and validation loss during multiple epochs of training, which can demonstrate that the model did not exhibit overfitting during the training process. As the number of training epochs increases, both the training loss (Train Loss) and the validation loss (Val Loss) show a decreasing trend, and the difference between the two is not significant. This indicates that the performance of the model on the training and validation sets is relatively consistent, and there is no situation where the training loss continues to decrease while the validation loss increases or stagnates, which is a positive signal of good model performance.

To sum up, the JKEM has demonstrated excellent performance in entity and relationship prediction tasks on the legal knowledge corpus. The model achieved high accuracy, recall, and F1 values in identifying various types of legal entities and relationships, particularly reaching near-perfect levels in the recognition of key information such as crime and defense. Although there are certain challenges in the recognition of complex texts, like case entities, overall, the JKEM model provides an efficient and accurate solution for processing legal texts. Future work could focus on further improving the model's

performance in specific scenarios such as complex case descriptions and exploring more strategies for integrating domain knowledge to enhance the model's generalization ability and prediction accuracy.
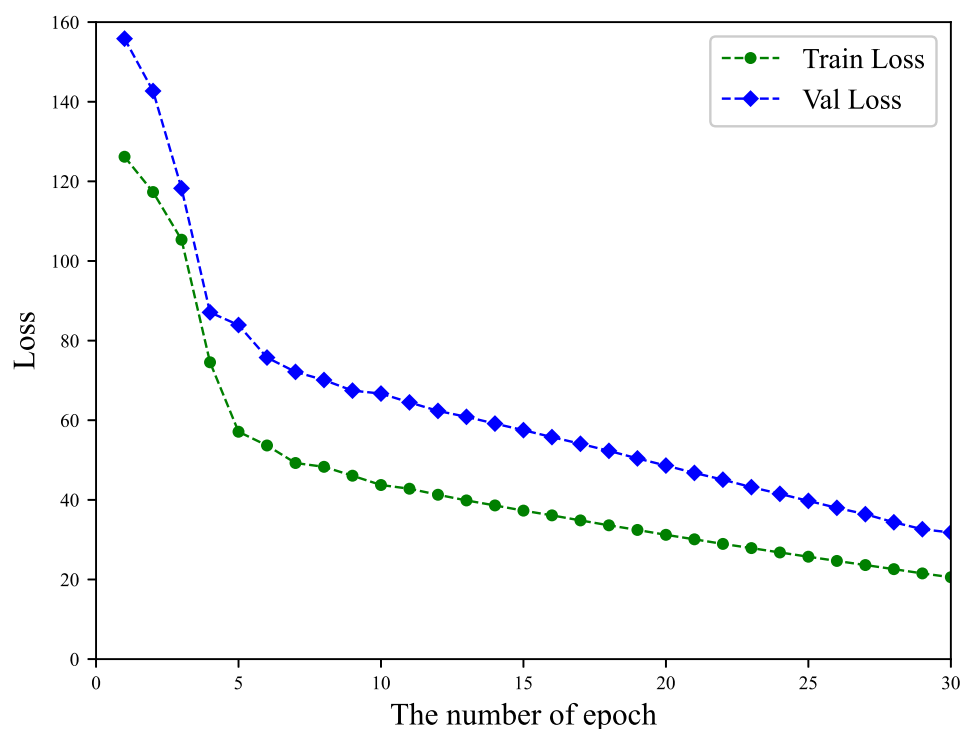

Figure 5. Changes in model training loss and validation loss during multiple epochs of training.

6. The Construction of CLKG

In the field of academic research, the effective storage and management of knowledge graphs play critical parts in supporting complex information analysis and knowledge discovery. This paper focuses on the storage mechanism of knowledge graphs and specifically selects Neo4j, a high-performance graph database system, as the core storage platform. As a widely used open-source graph database solution, Neo4j is not only renowned for its strong processing ability of graph data structures but also greatly promotes the collaboration efficiency between data scientists and developers by providing convenient interfaces with mainstream programming languages such as Python. In addition, Neo4j's support for a variety of graph mining algorithms lays a solid technical groundwork for the in-depth analysis and application of knowledge graphs.

In this study, we designed and implemented a process based on a knowledge extraction model, which successfully extracted 3480 pairs of high-quality entity-relationship triples from the source data. These triples constitute the basic data for constructing the knowledge graph in the legal field. Subsequently, the efficient import of these triple data and their structured storage in the graph database are realized through the seamless integration interface between Neo4j and Python. This process not only validates the efficiency of Neo4j in handling large-scale knowledge graph data but also demonstrates its potential in promoting the full-chain automation of data from extraction to storage.

To visually present the structure and content of the knowledge graph, the built-in visualization tools of Neo4j are further explored in this paper. However, due to the large scale and complexity of the overall knowledge graph, a direct display may be difficult to comprehensively and clearly reveal the fine-grained information of a specific domain (such as the specific criminal charges in the legal knowledge system). To this end, we adopted a focusing strategy and took "crimes endangering public safety" as an example, where Figure 6 was generated through the visualization interface of Neo4j.

As shown in Figure 6, the subcrimes of "crime endangering public security" are completely displayed, as well as the concepts, constitutive characteristics, judging standards, judicial interpretations, and other knowledge elements of various crimes. This figure accurately presents the entity association network around the crime, which effectively addresses the challenges faced by large-scale knowledge graphs in direct visualization and provides strong support for an in-depth understanding of the complex relationships in specific legal fields.
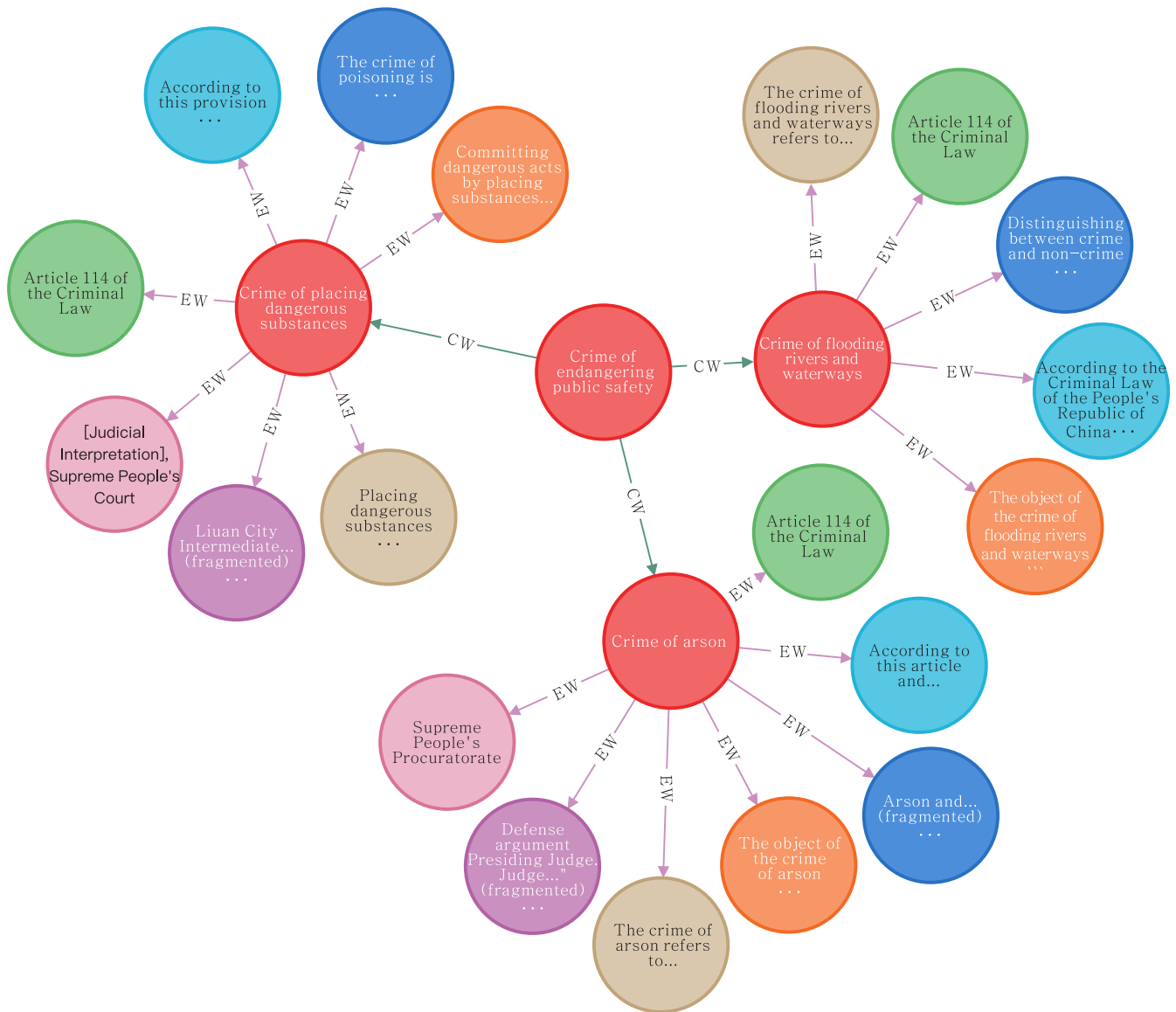
Figure 6. Sample of knowledge subgraph on crimes against public safety.

7. Conclusions and Future Work

A knowledge-enhanced LLM is developed in this paper for the construction of the CLKG in the context of in-depth exploration of the intrinsic correlation value of the massive datasets in the legal field. To this end, the prior knowledge is first seamlessly integrated into the LLM, and then a JKEM is developed by implementing a prefix fine-tuning strategy based on prior knowledge data. This fine-tuning strategy ingeniously embeds deep hints as continuous prefixes into the input sequences of each layer of the model while keeping the main parameters of the LLM frozen, which achieves customized enhancement of the LLM without manual feature engineering. Under the unsupervised or weakly supervised conditions, it achieves an accuracy of 90.92% and continuously shows excellent performance

in terms of accuracy, recall, and F1 score, significantly accelerating the efficiency and effectiveness of constructing the professional knowledge graph in the legal field.

Furthermore, this paper designs a knowledge graph architecture for legal knowledge, based on which the Chinese legal knowledge corpus is systematically collected and constructed from the authoritative text of the Chinese legal code and extensive internet resources. This corpus not only covers the original content of the legal provisions but also conducts multilevel and multidimensional depth analysis and expansion of the legal provisions through detailed annotations, professional interpretations, and rich case analysis. Thus, a comprehensive and in-depth legal knowledge system framework was constructed.

Ultimately, this paper successfully constructs a CLKG that elaborately depicts the knowledge details and structure in the legal field, laying a solid knowledge foundation and technical support for the subsequent development of the legal knowledge reasoning system based on the knowledge graph. The graph we constructed integrates 3480 knowledge triples composed of 9 core entities and 2 key relationships, providing strong data support for the intelligent application and decision support of legal knowledge.

While the present study has demonstrated the effectiveness of the proposed JKEM in constructing the CLKG and achieved promising results, there are several directions for future work to further advance the practical applications and theoretical foundations of this research.

First, we aim to expand on the practical implementations of the CLKG. Offering real-world examples of how the CLKG can be utilized in legal reasoning would significantly clarify its practical applications. For instance, integrating the CLKG with legal decision support systems could provide judges and legal practitioners with more comprehensive and accurate information, aiding in the interpretation of legal provisions and the prediction of judicial outcomes. By showcasing concrete use cases, we can better demonstrate the value of the CLKG in real-world legal scenarios.

Second, we intend to address any challenges encountered during the deployment of the CLKG. The process of deploying a knowledge graph in a complex legal environment is likely to face various obstacles, such as data privacy concerns, compatibility issues with existing legal systems, and the need for continuous updates to reflect the latest legal developments. By systematically analyzing and addressing these challenges, we can enhance the discussion of the practical feasibility and scalability of the CLKG.

Finally, we plan to explore how the model could be applied to different legal jurisdictions. The current study focuses on the construction of the CLKG based on Chinese legal provisions. However, the principles and methods employed in this research have the potential to be adapted and extended to other legal systems. By investigating the similarities and differences between various legal jurisdictions, we can develop strategies for customizing the CLKG to fit the specific requirements of different legal environments. This would not only broaden the applicability of the model but also facilitate cross-jurisdictional legal research and collaboration.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Yu, P.S. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 494–514. [CrossRef] [PubMed]
2. Kejriwal, M. Knowledge Graphs: A Practical Review of the Research Landscape. *Information* **2022**, *13*, 161. [CrossRef]
3. Liu, P.; Qian, L.; Zhao, X.; Tao, B. Joint Knowledge Graph and Large Language Model for Fault Diagnosis and Its Application in Aviation Assembly. *IEEE Trans. Ind. Inform.* **2024**, *20*, 8160–8169. [CrossRef]
4. Vrandečić, D.; Krötzsch, M. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
5. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; et al. DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. *Semant. Web* **2015**, *6*, 167–195. [CrossRef]
6. Liu, P.; Qian, L.; Zhao, X.; Tao, B. The Construction of Knowledge Graphs in the Aviation Assembly Domain Based on a Joint Knowledge Extraction Model. *IEEE Access* **2023**, *11*, 26483–26495. [CrossRef]
7. Hubauer, T.; Lamparter, S.; Haase, P.; Herzig, D.M. Use Cases of the Industrial Knowledge Graph at Siemens. In Proceedings of the Semantic Web—ISWC 2018, Cham, Switzerland, 8–12 October 2018; pp. 1–2.
8. Al-Moslmi, T.; Gallofré Ocaña, M.; Opdahl, A.L.; Veres, C. Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access* **2020**, *8*, 32862–32881. [CrossRef]
9. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [CrossRef]
10. Wu, T.; You, X.; Xian, X.; Pu, X.; Qiao, S.; Wang, C. Towards deep understanding of graph convolutional networks for relation extraction. *Data Knowl. Eng.* **2024**, *149*, 102265. [CrossRef]
11. Zheng, S.; Hao, Y.; Lu, D.; Bao, H.; Xu, J.; Hao, H.; Xu, B. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing* **2017**, *257*, 59–66. [CrossRef]
12. He, C.; Tan, T.P.; Zhang, X.; Xue, S. Knowledge-Enriched Multi-Cross Attention Network for Legal Judgment Prediction. *IEEE Access* **2023**, *11*, 87571–87582. [CrossRef]
13. Vuong, T.H.Y.; Hoang, M.Q.; Nguyen, T.M.; Nguyen, H.T.; Nguyen, H.T. Constructing a Knowledge Graph for Vietnamese Legal Cases with Heterogeneous Graphs. In Proceedings of the 2023 15th International Conference on Knowledge and Systems Engineering (KSE), Hanoi, Vietnam, 18–20 October 2023; pp. 1–6.
14. State Council Legislative Affairs Office (Compiler). *Criminal Law Code of the People's Republic of China: Annotated Edition (Fourth New Edition)*; Annotated Edition; China Legal Publishing House: Beijing, China, 2018.
15. Tagarelli, A.; Zumpano, E.; Anastasiu, D.C.; Calì, A.; Vossen, G. Managing, Mining and Learning in the Legal Data Domain. *Inf. Syst.* **2022**, *106*, 101981. [CrossRef]
16. Re, R.M.; Solow-Niederman, A. Developing Artificially Intelligent Justice. *Stanf. Technol. Law Rev.* **2019**, *22*, 242.
17. Remus, D.; Levy, F.S. Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law. *Georget. J. Leg. Ethics* **2015**, *30*, 501. [CrossRef]
18. Yao, S.; Ke, Q.; Wang, Q.; Li, K.; Hu, J. Lawyer GPT: A Legal Large Language Model with Enhanced Domain Knowledge and Reasoning Capabilities. In Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering (RAIIE '24), Singapore, 5–7 July 2024; pp. 108–112.
19. Savelka, J. Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts. In Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL '23), Braga, Portugal, 19–23 June 2023; pp. 447–451.
20. Ammar, A.; Koubaa, A.; Benjdira, B.; Nacar, O.; Sibaee, S. Prediction of Arabic Legal Rulings Using Large Language Models. *Electronics* **2024**, *13*, 764. [CrossRef]
21. Licari, D.; Bushipaka, P.; Marino, G.; Comandé, G.; Cucinotta, T. Legal Holding Extraction from Italian Case Documents using Italian-LEGAL-BERT Text Summarization. In Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL '23), Braga, Portugal, 19–23 June 2023; pp. 148–156.
22. Moreno Schneider, J.; Rehm, G.; Montiel-Ponsoda, E.; Rodríguez-Doncel, V.; Martín-Chozas, P.; Navas-Loro, M.; Kaltenböck, M.; Revenko, A.; Karampatakis, S.; Sageder, C.; et al. Lynx: A knowledge-based AI service platform for content processing, enrichment and analysis for the legal domain. *Inf. Syst.* **2022**, *106*, 101966. [CrossRef]
23. Tong, S.; Yuan, J.; Zhang, P.; Li, L. Legal Judgment Prediction via graph boosting with constraints. *Inf. Process. Manag.* **2024**, *61*, 103663. [CrossRef]
24. Bi, S.; Ali, Z.; Wu, T.; Qi, G. Knowledge-enhanced model with dual-graph interaction for confusing legal charge prediction. *Expert Syst. Appl.* **2024**, *249*, 123626. [CrossRef]
25. Zou, L.; Huang, R.; Wang, H.; Yu, J.X.; He, W.; Zhao, D. Natural language question answering over RDF: A graph data driven approach. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, UT, USA, 22–27 June 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 313–324.
26. Chen, J.; Teng, C. Joint entity and relation extraction model based on reinforcement learning. *J. Comput. Appl.* **2019**, *39*, 1918–1924.
27. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3580–3599. [CrossRef]
28. Yang, R.; Zhu, J.; Man, J.; Fang, L.; Zhou, Y. Enhancing text-based knowledge graph completion with zero-shot large language models: A focus on semantic enhancement. *Knowl.-Based Syst.* **2024**, *300*, 112155. [CrossRef]

29. Kumar, A.; Pandey, A.; Gadia, R.; Mishra, M. Building Knowledge Graph using Pre-trained Language Model for Learning Entity-aware Relationships. In Proceedings of the 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2–4 October 2020; pp. 310–315.

30. Zhang, Z.; Liu, X.; Zhang, Y.; Su, Q.; Sun, X.; He, B. Pretrain-KGE: Learning Knowledge Representation from Pretrained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 16 November 2020; pp. 259–266.

31. Zhang, J.; Chen, B.; Zhang, L.; Ke, X.; Ding, H. Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open* **2021**, *2*, 14–35. [CrossRef]

32. Abu-Salih, B. Domain-specific knowledge graphs: A survey. *J. Netw. Comput. Appl.* **2021**, *185*, 103076. [CrossRef]

33. Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Yang, B.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; et al. Never-ending learning. *Commun. ACM* **2018**, *61*, 103–115. [CrossRef]

34. Cadeddu, A.; Chessa, A.; De Leo, V.; Fenu, G.; Motta, E.; Osborne, F.; Reforgiato Recupero, D.; Salatino, A.; Secchi, L. Optimizing Tourism Accommodation Offers by Integrating Language Models and Knowledge Graph Technologies. *Information* **2024**, *15*, 398. [CrossRef]

35. Nakayama, H.; Kubo, T.; Kamura, J.; Taniguchi, Y.; Liang, X. Doccano: Text Annotation Tool for Human. Available online: https://github.com/doccano/doccano (accessed on 1 May 2020).

36. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 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), Online, 1–6 August 2021; pp. 4582–4597.

37. Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; Tang, J. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 320–335.

38. Goyal, A.; Gupta, V.; Kumar, M. Recent Named Entity Recognition and Classification techniques: A systematic review. *Comput. Sci. Rev.* **2018**, *29*, 21–43. [CrossRef]

39. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. In Proceedings of the NAACL, San Diego, CA, USA, 16 June 2016; pp. 260–270.

40. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.

41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Red Hook, NY, USA, 4–9 December 2017; pp. 6000–6010.