JurisGraph Insight Engine 1.0v: A Legal Question Answering System Based on Large Language Models and Knowledge Graphs

Haiguang Zhang MAXWELLZHG@GMAIL.COM
The Chinese University of Hong Kong, Shenzhen, Guangdong, China

Editors: Hung-yi Lee and Tongliang Liu

Abstract

The extraction and effective utilization of judicial data remains a major challenge in the legal domain. There is a growing mismatch between the public's demand for accessible legal services and the high cost and complexity of legal consultations, which also affects the efficiency of legal professionals when handling case inquiries. Traditional keyword-based search methods lack professionalism, interpretability, and scalability. In this paper, we propose JurisGraph Insight Engine 1.0v, an intelligent legal question answering (QA) system that integrates large language models (LLMs) and domain-specific knowledge graphs. We first construct a comprehensive Criminal Law Knowledge Graph (CLKG) containing 483 types of criminal offenses, and develop two unified heterogeneous subgraphs for theft and drug-related cases. Then, we fine-tune a domain-specific legal LLM, LawM, using a curated corpus of over 280,000 Chinese legal records covering multiple legal Natural Language Processing (NLP) tasks. Finally, we design and implement a QA system that leverages both the knowledge graph and LawM to deliver accurate and interpretable answers to legal questions. Experimental results show that our system achieves 95% accuracy, effectively lowering the barrier to legal knowledge access for the general public while improving decision efficiency for legal practitioners.

Keywords: Legal Question Answering; Knowledge Graph; Large Language Models; Criminal Law; Legal AI

1. Introduction

In an era of increasingly mature smart judiciary systems, ensuring accessible legal knowledge and enhancing legal literacy have attracted growing public attention. Rapid advances in the internet and AI technologies are driving society's transition from the information age to the intelligent age. The recent release of GPT-4 (OpenAI, 2023) has garnered widespread global attention, driving the development of LLMs with tens of billions of parameters and hundreds of billions of tokens. These LLMs, including PaLM (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023), and GLM (Zeng et al., 2022), have shown exceptional performance across various natural language processing tasks, achieving new state-of-the-art results. Compared to earlier pre-trained language models such as BERT (Devlin et al., 2018), LLMs exhibit superior text generation and comprehension capabilities. Leveraging these advantages, LLMs have been explored in various fields, including law (Cui et al., 2023a), education (Kasneci et al., 2023), finance (Wu et al., 2023), and biomedical and health sciences (Thirunavukarasu et al., 2023).

In the legal domain, especially criminal law, laypeople often face difficulties in extracting accurate legal information from vast and complex online content due to limited professional background. Meanwhile, emerging AI-powered legal assistants heavily rely on well-constructed knowledge graphs to interpret users' natural language queries. Legal entities and institutions have begun to construct domain-specific knowledge graphs based on statutes and case law. However, most existing legal knowledge graphs are stored in RDF format, which poses challenges for non-technical users to navigate and utilize effectively. Moreover, the surge in legal documents and fragmented data presents new challenges. Free legal aid resources remain limited and not easily accessible for the public, while judicial professionals must often sift through large volumes of documentation to identify criminal records, a time-consuming and labor-intensive task.

To overcome these challenges, we propose an integrated legal question-answering system that combines a large-scale Criminal Law Knowledge Graph with a fine-tuned legal language model, LawM. Unlike traditional keyword-based search engines that return lengthy and often irrelevant results, our system retrieves structured legal knowledge and generates precise, context-aware answers tailored for both legal professionals and non-experts. It supports intuitive natural language queries, graphical answer visualization, and multi-type case inquiries, all within a privacy-preserving framework. The main contributions of this paper are as follows:

- We construct a large-scale criminal law knowledge graph, integrating heterogeneous case types with unified ontologies.
- We develop LawM, a legal-domain large language model fine-tuned on 280K Chinese legal records, outperforming several baselines.
- We build JurisGraph Insight Engine, a multi-functional QA system enabling intuitive, accurate legal consultation and case search for professionals and non-professionals alike.
- We propose a domain-specific, multi-layer integration framework combining hierarchical knowledge graphs (statutory and case law), sensitivity-aware question routing, and multimodal explainability, delivering a practical and interpretable legal AI system tailored for real-world Chinese criminal law scenarios.

2. Related Work

2.1. Legal Large Language Model

In the legal domain, LLMs have emerged as promising tools for delivering personalized legal assistance. However, legal texts differ substantially from general-domain texts in several key aspects: they involve complex legal terminology, ambiguous abbreviations, intricate syntactic structures, and highly specialized vocabulary. These unique characteristics present significant challenges for general-purpose LLMs, which are typically trained on open-domain corpora. To address this gap, several legal-domain LLMs have been developed. For example, ChatLaw (Cui et al., 2023a) applies fine-tuning techniques based on Ziya-LLaMA and LoRA, achieving strong performance in legal QA. Disc-LawLLM (Yue et al., 2023) proposes a Chinese-language intelligent legal system powered by LLMs, designed to serve a wide range of user needs. Similarly, LawyerLLaMA (Huang et al., 2023)

fine-tunes the LLaMA model using legal documents, with a primary focus on improving performance in legal QA and dialogue-oriented tasks. These legal LLMs are generally adapted from open-source general models such as LLaMA (Touvron et al., 2023) and GLM (Zeng et al., 2022), and are fine-tuned on domain-specific QA and dialogue datasets. Despite advances, the capabilities of legal LLMs on tasks such as statute retrieval, case classification, and legal reasoning are still insufficiently studied, underscoring the need for comprehensive investigation.

2.2. Knowledge Graph-Based Question Answering (KGQA)

Knowledge graphs represent semantic relationships between entities using structured triples (Pan et al., 2017), and have been widely applied in domains such as commerce (Li et al., 2020a), healthcare (Li et al., 2020b), agriculture (Qin and Yao, 2021), and law (Filtz, 2017). Most KGQA methods integrate deep learning or embed both questions and knowledge triples into a shared vector space for answer retrieval (Dong et al., 2015). Early works explored semantic parsing (Reddy et al., 2014), feature engineering (Yao and Van Durme, 2014), and attention mechanisms (Qu et al., 2018) to enhance reasoning over graphs. Answer generation approaches fall into two main categories: template-based methods (Cui et al., 2019), which require manual template construction, and semantic parsing-based methods (Wang et al., 2020), which map questions into executable logical forms. While recent advances have improved performance, challenges remain in handling complex queries and ensuring domain adaptability.

2.3. Deep Learning-Based Knowledge Graph Question Answering

Deep learning has been widely adopted in KGQA, primarily in two ways: enhancing individual components of traditional QA pipelines and enabling end-to-end learning. In the former, deep models improve tasks such as entity recognition, relation classification, and logical form generation (Hua et al., 2020). For example, STAGG (Yih et al., 2015) incrementally builds query graphs to reduce semantic parsing complexity. In the latter, end-to-end models directly map questions to answers via embedding-based ranking and retrieval. Bordes et al.(Bordes et al., 2014) proposed embedding both question phrases and candidate entities into a shared space, while Dong et al.(Dong et al., 2015) used CNNs to encode questions, answer types, and relations. Subsequent work further refined retrieval performance using neural ranking methods (Xiong et al., 2019).

3. Methodology

This paper presents a comprehensive legal natural language processing framework, as illustrated in Figure 1. The framework comprises three core components: (1) the construction of a Criminal Law Knowledge Graph (CLKG), a subgraph of LawKG, and case-specific knowledge graphs for drug-related and theft cases, namely the Criminal Drug Knowledge Graph (CDKG) and the Criminal Theft Knowledge Graph (CTKG); (2) the design and training of LawM; and (3) a knowledge graph-based QA system, JurisGraph Insight Engine. By integrating these knowledge graphs, the proposed framework enables accurate legal information retrieval and QA tailored to criminal law scenarios.

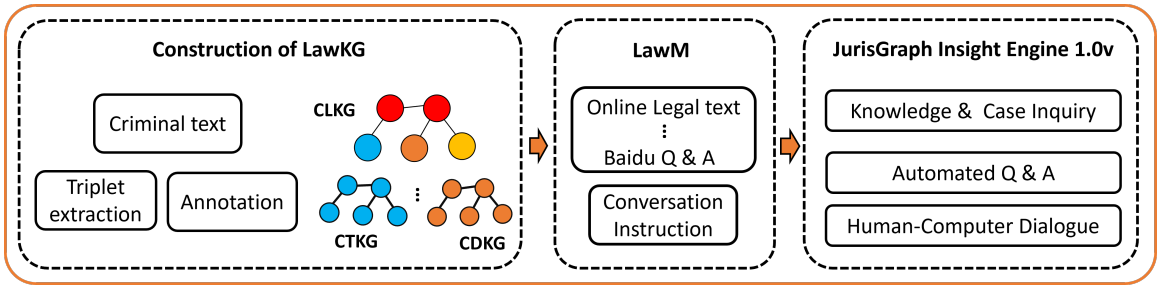
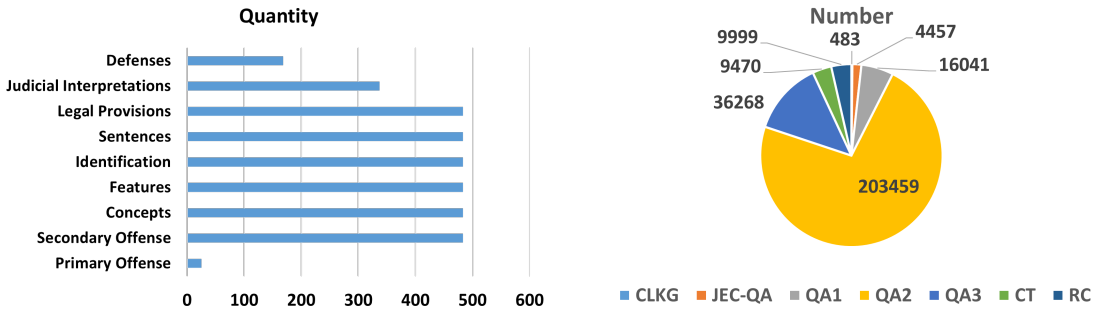


Figure 1: Architecture of JurisGraph Insight Engine 1.0v

3.1. Construction of CLKG

This study adopts a bottom-up design strategy for constructing the legal knowledge graph. The overall architecture is logically divided into two layers: the schema layer and the data layer. The schema layer serves as a foundational component, responsible for building an ontology repository that defines key legal concepts and their interrelationships. Ontological elements are represented using triples, such as entity–relation–entity structures or entity–attribute–value combinations. The data layer is dedicated to storing these structured triples. Data related to criminal law offenses are collected from national legal databases and publicly available legal websites. After systematic addition and refinement, the knowledge graph includes 483 distinct offense entries, as summarized in Figure 2(a).



(a) Statistics of Entities and Concepts in the CLKG

(b) Overview of Dataset Composition

Figure 2: Statistics of CLKG and Dataset Composition

In this figure, Primary Offenses refer to higher-level offense categories, serving as classifications for Secondary Offenses. For example, the secondary offense "Crime of Obstructing Drug Administration" falls under the Primary Offense "Crime of Undermining the Socialist Market Economy Order." Concepts denote the formal definitions of these offenses. Features describe attributes that characterize each crime, while Identification delineates the boundaries that distinguish one offense from others. Sentences represent the punishments imposed on offenders, while Legal provisions specify the statutory basis for these penalties. Judicial interpretations consist of official explanations issued by the highest national judicial authorities regarding specific legal issues encountered during law enforcement. Defense records contain defense statements from relevant past cases.

This paper details the design of entity and relationship types, as illustrated in Figure 3. Guided by legal experts and with the help of volunteers who carefully annotated the data, including key elements such as the subject, the subjective and objective aspects, and the object of the crime, we identified 13 entity types and 12 relationship types, providing a solid foundation for building the legal knowledge graph. To ensure data quality, rigorous verification and thorough data cleaning processes were conducted to eliminate redundancy. The resulting data was stored in a Neo4j graph database, which models structured data as a network rather than in traditional tabular form, thus aligning with knowledge graph characteristics.

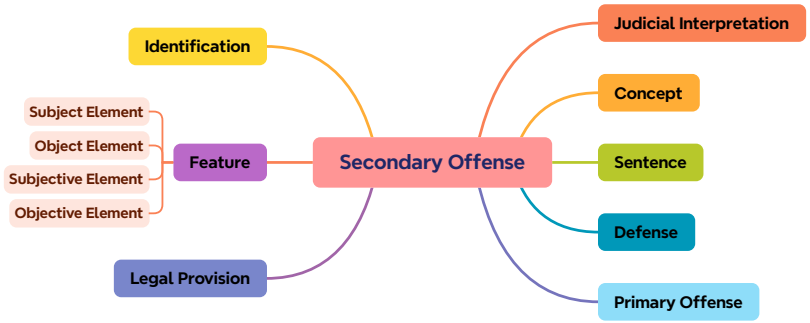


Figure 3: System Architecture of CLKG

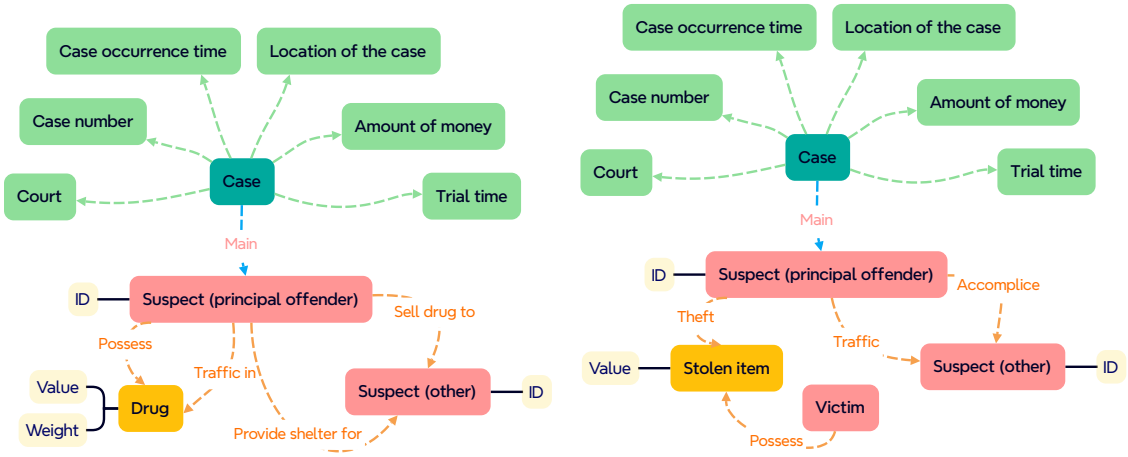
3.2. Construction of a Unified Knowledge Graph for Heterogeneous Crime Cases

This study focuses on constructing a unified knowledge graph for heterogeneous crime cases, with the goals of systematically organizing case information, preventing the omission of critical criminal details, and ensuring data accuracy through real-time or periodic updates. The datasets for theft and drug trafficking cases are primarily sourced from the China Judgments Online database and from (Zhang et al., 2024), which provide publicly accessible indictments, verdicts, and related documents containing rich case information.

To ensure comprehensive coverage and enhance the extensibility of the heterogeneous crime knowledge graph, the ontology design has been carefully refined. A set of entity types has been defined, with relationship types tailored to different case categories. Figures 4(a) and 4(b) depict the detailed entity and relationship schemas, laying a robust foundation for constructing a precise and comprehensive legal knowledge graph, which is stored using the Neo4j graph database. Following construction, the accuracy of entities and the validity of relationships were thoroughly verified to ensure that all connections faithfully represent real-world associations within the domain. This validation process aims to enhance the overall quality and practical value of the knowledge graph.

3.3. Law Large Model (LawM)

To evaluate the effectiveness of fine-tuned LLMs in legal NLP tasks, this paper proposes LawM, a legal domain-specific LLM designed for diverse downstream applications. The model is trained on a diverse legal text mining dataset covering multiple task types and



(*a*) CDKG (*b*) CTKG

Figure 4: Design of CDKG and CTKG

demonstrates versatility across various legal challenges. While generative LLMs can handle multitasking through supervised fine-tuning, non-generative tasks such as information extraction may still favor smaller, discriminative models.

Most existing legal LLMs focus on QA and dialogue, but often underperform in specific legal subdomains. Models such as LexiLaw (Du et al., 2021), LawyerLLaMA (Huang et al., 2023), and Chat-law (Cui et al., 2023a) demonstrate strong performance in their respective domains. Distinct from these, LawM emphasizes real-world applicability and systematically explores the capabilities of large models within the legal domain.

LawM is trained in two stages: an instruction fine-tuning phase using general instruction datasets for broad task comprehension, followed by a domain-specific phase leveraging professional legal corpora to impart specialized expertise. This training pipeline can be summarized as follows:

$$\phi_1 = \arg\min_{\phi} \mathcal{L}_{\text{general}}(\phi \mid \mathcal{D}_{\text{general}}), \quad \phi_2 = \arg\min_{\phi} \mathcal{L}_{\text{legal}}(\phi \mid \mathcal{D}_{\text{legal}}) \tag{1}$$

Here, ϕ denotes the model parameters; $\mathcal{L}_{\text{general}}$ and $\mathcal{L}$legal represent the loss functions on the general instruction dataset $\mathcal{D}_{\text{general}}$ and the legal domain dataset $\mathcal{D}_{\text{legal}}$, respectively. Combining these two training stages, our approach can be summarized as follows:

$$\phi^* = \arg\min_{\phi} \left(\mathcal{L}_{\text{general}}(\phi \mid \mathcal{D}_{\text{general}}) + \mathcal{L}_{\text{legal}}(\phi \mid \mathcal{D}_{\text{legal}})\right) \tag{2}$$

Here, ϕ^* denotes the final model parameters obtained after the two-stage training process.

3.4. Question Answering System

In this study, we employ web scraping techniques to gather relevant legal data and store high-quality textual resources. We integrate the knowledge graph, search architecture, and question-answering modules into a unified system that supports retrieval and QA functionalities within a private domain. A question classification module is incorporated to enhance

classification accuracy, better capture user intent, and provide preliminary answer predictions, thereby reducing subsequent processing costs and computational load on the model. The system first performs keyword extraction on user queries, then leverages deep learning methods combined with the knowledge graph to generate accurate answers or retrieve them from the knowledge base. Designed with multiple functionalities (see Figure 5), the system effectively addresses diverse user needs while maintaining overall robustness and integrity.

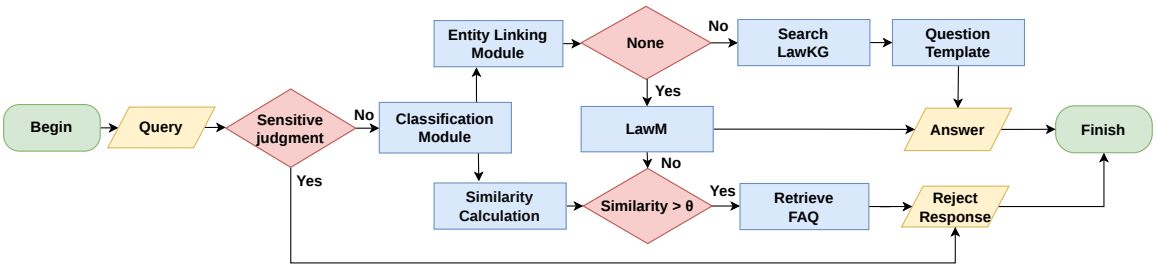


Figure 5: Flowchart of JurisGraph Insight Engine 1.0v

The system classifies and identifies entities within user queries to determine the question category. For questions involving sensitive categories, the system provides appropriate error feedback or prompts the user to rephrase their query.

$$C = f_{classify}(Q) \tag{3}$$

$$Judge = \begin{cases} Error & if C \in SensitiveCategories, \\ Proceedwith & otherwise. \end{cases} \tag{4}$$

The system routes questions to relevant processing modules and employs specific query paths to extract information from the CLKG, a subgraph of LawKG. Retrieved answers are presented alongside corresponding subgraphs for context.

$$Answer, Subgraph = f_{query}(Q, CLKG) \tag{5}$$

To provide legal advisory services, the system integrates a Frequently Asked Questions (FAQs) module that computes similarity between the user's question and database entries. If similarity exceeds a threshold θ, the corresponding database answer is returned; otherwise, LawM generates a response.

$$Answer = \begin{cases} Retrieve(Q_i) & if Similarity(Q, Q_i) > \theta, \\ GenerateAnswer(Q, LawM) & otherwise. \end{cases} \tag{6}$$

Legal practitioners can retrieve relevant case details by submitting perpetrator identity information (ID). The system queries the knowledge graph and returns case data along with associated graph information upon successful search.

$$Answer, Subgraph = f_{caseQuery}(ID, CLKG) \tag{7}$$

4. Experiments

4.1. Data Processing

As shown in Figure 2(*b*), the training data are collected from open-source legal datasets. Here, QA denotes Legal Question Answering, CT refers to Cases and Trials, and RC stands for CAIL Reading Comprehension. To better align with large language model training, judicial examination data are filtered and converted into single-turn dialogue pairs by matching questions with correct answers. Specifically, CLKG and RC represent multi-turn dialogue datasets, whereas the others are designed for single-turn interactions. After data cleaning, the final dataset contains approximately 280,000 records. The dataset is then split into training and testing sets at a ratio of 99:1.

For evaluation purposes, we also prepare task-specific datasets. For Named Entity Recognition (NER), we use a legal dataset annotated from Chinese Criminal Law statutes and case descriptions, with all entities double-annotated by legal experts (Cohen's Kappa = 0.82). For charge classification, we construct a dataset from Chinese criminal case documents, labeling each case segment with its corresponding statutory charge. Both datasets are used to assess LawM on sequence labeling and classification tasks.

4.2. Knowledge Graphs

The constructed criminal knowledge graphs provide a structured and comprehensive representation of the complex relationships within drug-related and theft-related criminal activities. The CDKG captures detailed information on illicit substances, distribution networks, associated criminal organizations, trafficking routes, smuggling methods, and money laundering schemes. Meanwhile, the CTKG encompasses various elements of theft cases, including target demographics, modus operandi, disposal channels for stolen goods, and underground market connections.

By visualizing these relationships, the knowledge graphs reveal the underlying mechanisms driving criminal behaviors and support informed decision-making for law enforcement and research purposes. Furthermore, a comparison of entity and relationship distributions across CDKG and CTKG highlights the differences in structural complexity and information density between drug and theft cases. Detailed statistics, including node counts, edge types, and attribute distributions, are provided in Tables in the appendix. These tables offer a comprehensive overview of the graph structures, allowing readers to better understand the scope and richness of the data captured.

4.3. System Implementation and Testing

We fine-tuned Qwen-7B (Bai et al., 2023) using QLoRA on a single NVIDIA RTX 3090 (24GB). Training was carried out with a learning rate 2×10^{-4}, batch size 4, maximum sequence length 1024, LoRA rank 16, $\alpha = 16$, dropout 0.05, and 700 warm-up steps for one epoch (approximately 25 hours).

For reproducibility, maintenance, and future development, we report the software and hardware environment as follows: the system runs on Microsoft Windows 10; the knowledge graph database uses Neo4j version 4.0.7; the distributed search engine is Elasticsearch version 6.6.0; development is conducted with VSCode 1.86; the front end framework is Flask

1.1.2; the programming language is Python 3.9.7; and key dependencies include Py2neo 2021.2.4 and Selenium 4.1.0. The system interface consists of two main sections: a dialogue box on the left that records conversation history and allows user query input, and a knowledge graph display panel on the right that retrieves and visualizes relevant subgraphs from the database based on user queries. Responses generated by the system are displayed in the left side dialogue box.

Functional testing confirmed that the system meets requirements, including legal QA, legal knowledge queries, and case search capabilities. A suite of test cases was designed to verify the system's ability to receive and parse user submitted questions, identify sensitive content, retrieve relevant knowledge graph nodes, and conduct dialogue based QA. For each feature, 10 random test samples were executed. The testing procedure included the user entering a query in the dialogue box, clicking the send button, the system entering a processing state, performing sensitivity checks, and returning the appropriate response. The expected behavior was to reject sensitive questions and successfully parse and answer non-sensitive ones. The system's average response time measures the typical duration needed to process and answer user queries across different functionalities, reflecting the system's responsiveness. The BLEU score, commonly used in machine translation and text generation evaluation, quantifies the similarity between generated and reference texts. This paper evaluates the system both in terms of the quality of generated content and response time, with results summarized in Table 1. The test outcomes demonstrate that the system performs as expected, with accurate question parsing, normal feature operation, and strong security and reliability characteristics.

Table 1: System Evaluation Results

Metric	Avg Time	Accuracy	BLEU
Value	1.41s ± 0.15s	95% ± 2.1%	63% ± 3.6%

4.4. Experimental Results

To comprehensively evaluate the generative performance of our model, we employed multiple metrics,Accuracy (Acc), Completeness (Cpt), and Clarity (Cla) scored on a 1–5 scale by both ChatGPT and 10 human annotators. The evaluation was conducted on original datasets, including the benchmark JEC-QA dataset (Zhong et al., 2020; Zhang et al., 2022), which focuses on Chinese legal QA with judicial and bar exam-style queries. Detailed scoring rubrics with concrete examples were provided to all annotators to ensure consistent and objective assessments (e.g., a score of 5 indicates fully correct and clear answers, while 3 indicates partially correct or unclear responses). Inter-annotator agreement was measured on a subset of 150 samples annotated by five independent annotators, with Fleiss' Kappa values showing substantial agreement across metrics: Acc ($\kappa = 0.79$), Cpt ($\kappa = 0.75$), and Cla ($\kappa = 0.77$), supporting the reliability of the human evaluation process, as shown in Table 2.

For evaluation, we compared our model with several strong baselines, including ChatGLM (Du et al., 2021), Baichuan-Chat (Baichuan, 2023), Chinese-Alpaca-2 (Cui et al.,

Table 2: Evaluation Results of the Large Language Models. Boldface indicates the highest score. QA: LegalQA; CT: Cases and Trials; RC: CAIL Reading Comprehension.

Dataset	LawM				Human				Raw Data			
	Acc	Cpt	Cla	Avg	Acc	Cpt	Cla	Avg	Acc	Cpt	Cla	Avg
QA1	4.42±0.13	3.58±0.10	3.86±0.09	3.95±0.12	4.38±0.11	3.67±0.08	3.88±0.10	3.98±0.13	4.18±0.10	3.44±0.09	3.53±0.11	3.72±0.12
QA2	4.62±0.12	3.89±0.10	4.04±0.11	4.18±0.09	4.53±0.13	3.79±0.08	4.25±0.12	4.19±0.12	4.37±0.11	4.06±0.10	3.97±0.09	4.13±0.11
QA3	4.55±0.14	3.84±0.09	4.00±0.10	4.13±0.10	**4.86±0.10**	3.94±0.08	**4.30±0.11**	4.37±0.12	4.46±0.10	3.94±0.09	3.96±0.08	4.12±0.10
JEC-QA	**4.95±0.07**	**4.88±0.06**	**4.84±0.08**	**4.89±0.07**	4.67±0.09	4.30±0.08	**4.79±0.06**	**4.59±0.07**	4.83±0.06	4.76±0.05	**4.72±0.07**	4.77±0.07
CT	4.25±0.12	3.79±0.11	3.66±0.10	3.92±0.11	4.35±0.10	3.84±0.09	3.50±0.08	3.90±0.09	4.39±0.10	4.22±0.08	4.17±0.07	4.26±0.09
RC	4.12±0.11	3.54±0.10	3.72±0.09	3.79±0.10	4.34±0.12	3.35±0.09	3.96±0.10	3.88±0.11	**4.88±0.08**	4.50±0.06	4.53±0.07	4.63±0.07
CLKG	4.33±0.10	3.87±0.09	3.95±0.10	4.05±0.10	4.54±0.11	4.27±0.10	4.03±0.08	4.28±0.09	4.82±0.07	**4.83±0.06**	4.67±0.07	**4.78±0.07**

2023b), GPT-3.5-turbo (OpenAI, 2022), LexiLaw (University, 2023b), LawGPT (University, 2023a), LawyerLLaMa (Huang et al., 2023), ChatLaw (Cui et al., 2023a), DISC-LawLLM (Yue et al., 2023), InternLM-Law (Fei et al., 2025), and DeepSeek (DeepSeek-AI et al., 2025). These models cover both general-purpose LLMs and domain-specific legal LLMs, providing a comprehensive benchmark for our study.

Table 3 presents comparative results against several open-source large language models, illustrating our model's competitive performance. Additionally, despite having only 7 billion parameters, our model achieves strong results on the NJE dataset's PAE and CPA multiple-choice questions, ranking second and third respectively. The average score of 29.92 places it just behind larger models such as Baichuan, GPT-3.5, and DISC, reflecting its robust capabilities in text generation and legal question comprehension. Compared to recent systems such as ChatLaw, DISC-LawLLM, and LawGPT, JurisGraph emphasizes real-world deployability, legal interpretability, and privacy preservation. The proposed framework distinguishes itself through the integration of structured legal subgraphs (CTKG, CDKG) aligned with criminal code hierarchies, the design of a privacy-preserving and rejectable question–answering interface for handling ambiguous or sensitive queries, and a multimodal explainability mechanism that couples natural language responses with interactive legal subgraph visualizations, thereby enhancing transparency and auditability for legal practitioners.

Table 3: Evaluation Results of LawM. † denotes results directly quoted from DISC-LawLLM (Yue et al., 2023). Boldface indicates the highest score. S and M are shorthand of single-answer and multiple answers, respectively. National Judicial Examination (NJE), Patent Agent Examination (PAE), Certified Public Accountant Qualification Examination (CPA), Unified National Graduate Entrance Examination (UNGEE), Public Institutions and Functionary Examination (PFE) and Question Bank of Legal Basic Knowledge (LBK).

Model	Size	Hard						Normal		Easy		Avg
		NJE		PAE		CPA		Ungee		PFE	LBK	
		S	M	S	M	S	M	S	M	S	S	
GPT-3.5-turbo†	175B	36.50	10.58	37.29	17.03	**42.13**	**21.67**	**51.25**	**28.74**	53.53	54.18	34.10
Lawyer LLaMa†	13B	35.75	5.62	32.20	6.52	29.95	13.33	32.50	14.94	39.41	39.64	25.05
ChatLaw†	13B	27.56	7.99	31.36	9.42	35.53	11.67	35.62	17.24	42.35	41.09	25.20
DISC-LawLLM†	13B	**42.09**	**19.87**	**40.68**	**18.48**	39.59	19.17	50.94	25.29	**57.06**	**54.91**	**37.10**
Baichuan-Chat†	13B	31.47	10.15	29.66	8.70	35.53	19.17	50.00	27.59	53.12	53.45	30.78
Chinese-Alpaca-2 †	13B	25.70	10.15	30.51	11.59	32.99	19.17	40.94	21.84	44.12	43.27	26.73
DeepSeek	13B	37.12 ± 0.90	15.38±1.00	36.45±1.10	14.27±0.80	38.63±0.70	18.54±1.00	48.19±1.20	23.41±1.00	52.08±0.80	50.76±0.90	33.48±0.94
ChatGLM†	6B	31.66	1.08	27.97	2.90	37.06	13.33	39.69	20.69	37.65	42.91	24.66
LexiLaw†	6B	20.11	7.56	23.73	10.14	24.87	19.17	31.56	16.09	31.76	40.36	21.50
LawGPT †	7B	22.91	6.26	31.36	7.61	25.38	16.67	30.31	13.79	34.71	29.09	20.60
InternLM-Law	7B	26.13±0.67	7.12±0.91	27.88±1.15	9.08±0.82	27.05±1.01	14.92±0.76	32.27±1.10	14.41±0.68	35.28±0.89	31.15±0.55	22.53±0.85
LawM	7B	**37.57±1.32**	**15.22±0.96**	**30.81±1.11**	**16.20±1.05**	**39.43±0.94**	**20.14±0.89**	**33.96±1.08**	**26.84±1.13**	**48.76±1.25**	**30.24±1.07**	**29.92±1.08**

Table 4 reports the NER performance on our annotated legal dataset. LawM achieves an F1-score of 86.8%, slightly below BiLSTM-CRF (88.1%) but substantially higher than SVM (81.5%), demonstrating robust sequence labeling capability for legal entities. For charge classification (Table 5), LawM attains 84.7% accuracy, surpassing the SVM baseline at 79.5%, indicating effective semantic comprehension of legal texts. Overall, these results highlight LawM's versatility across both generative and discriminative legal NLP tasks, confirming its applicability in diverse real-world legal scenarios.

Table 4: NER performance comparison (%)

Model	Precision	Recall	F1-score
SVM	82.4	80.7	81.5
BiLSTM-CRF	88.9	87.3	88.1
LawM	87.1	86.5	86.8

Table 5: Charge Classification accuracy comparison (%)

Model	Accuracy
SVM	79.5
LawM	84.7

To assess robustness, we evaluated performance under varying train–test splits and across question types of differing complexity. As shown in Table 6, LawM maintains stable accuracy across splits, with F1-scores of 86.4%, 84.7%, and 82.3% for 99:1, 90:10, and 80:20 splits, respectively, indicating only minor declines under more balanced splits. Then, Table 7 presents performance by question type under the 80:20 split: fact-based queries achieve 84.6% F1, statute-based queries 81.6%, and multi-hop reasoning queries 78.0%, demonstrating modest degradation with increasing reasoning complexity.

We further benchmarked QA performance on 3,000 samples spanning theft, drug-related, and general crime domains. As shown in Table 8, the integrated model ("Ours") outperforms all baselines, achieving 93.7% accuracy, 0.64 BLEU, 89.5% statute precision, and 98.5% answerable rate. In comparison, LawM alone achieves 83.4% accuracy and 65.2% statute precision, KG-only achieves 78.0% accuracy with higher statute precision (91.3%) but lower answerable rate (82.7%), and Qwen-7B attains 71.8% accuracy and 42.7% statute precision. These results indicate that combining LawM with structured knowledge graphs synergistically enhances both fluent response generation and accurate statute retrieval, delivering consistent gains across all metrics.

A detailed breakdown by QA task is shown in Table 9. Our model achieves 95% accuracy in statute retrieval, substantially outperforming the baselines. Multi-hop reasoning accuracy reaches 89%, again clearly higher than LawM and KG-only. Factual recall 92% and ambiguous input handling 87% further demonstrate the robustness and consistent advantage of the integrated approach.

Table 6: Performance under different train-test splits (mean ± std %)

Train-Test Split	Precision	Recall	F1-score
99:1	87.3 ± 0.2	85.6 ± 0.4	86.4 ± 0.3
90:10	85.8 ± 0.3	83.7 ± 0.5	84.7 ± 0.4
80:20	83.2 ± 0.4	81.5 ± 0.5	82.3 ± 0.4

Table 7: Performance breakdown by question type (80:20 split), in %

Question Type	Precision	Recall	F1-score
Fact-based	85.1 ± 0.3	84.2 ± 0.3	84.6 ± 0.2
Statute-based	82.7 ± 0.4	80.5 ± 0.5	81.6 ± 0.3
Multi-hop Reasoning	79.3 ± 0.5	76.8 ± 0.6	78.0 ± 0.4

Table 8: Ablation results on QA performance (mean ± std %)

System	Accuracy	BLEU	Statute Precision	Answerable Rate
Ours	93.7 ± 0.8	0.64 ± 0.02	89.5 ± 1.1	98.5 ± 0.5
LawM	83.4 ± 1.2	0.59 ± 0.03	65.2 ± 2.4	99.1 ± 0.3
KG-only	78.0 ± 1.0	0.52 ± 0.04	91.3 ± 1.7	82.7 ± 1.5
Qwen-7B	71.8 ± 1.5	0.48 ± 0.03	42.7 ± 3.0	97.6 ± 0.6

Table 9: Answer accuracy by question type and system variant (mean ± std %)

Question Type	Ours	LawM	KG	Qwen-7B
Statute Retrieval	95 ± 1.2	72 ± 2.1	91 ± 1.5	55 ± 3.0
Multi-hop Reasoning	89 ± 1.5	76 ± 2.3	62 ± 2.0	48 ± 2.8
Factual Recall	92 ± 1.0	88 ± 1.8	80 ± 1.7	69 ± 2.4
Ambiguous Input	87 ± 1.3	82 ± 2.0	51 ± 3.2	45 ± 3.1

4.5. Domain Coverage and Fairness Analysis

To assess the fairness and reliability of our system across diverse legal scenarios, we categorize the evaluation set into four representative domains: Theft, Drug-related crimes, Fraud, and Traffic offenses. Among these, the Theft and Drug domains are supported by structured legal subgraphs (CTKG and CDKG), enabling more accurate and interpretable reasoning. In contrast, Fraud and Traffic are currently addressed using curated FAQ-style knowledge. Table 10 reports the domain specific performance in terms of (1) law citation accuracy, (2) answer completeness, and (3) lawful rejection rate. The results show that our subgraph-based reasoning yields consistently stronger performance, especially in citation accuracy and completeness. We plan to expand structured subgraph support to other domains in future work to ensure broader and fairer legal coverage. All reported values are averaged over 5 independent runs (for automatic evaluation) or 10 annotators (for human

Table 10: Per-domain evaluation results on legal QA performance (%)

Legal Domain	Law Citation Accuracy	Answer Completeness	Lawful Rejection Rate
Theft	92.1	88.5	12.4
Drug	94.0	91.0	9.8
Fraud	85.2	81.0	15.7
Traffic	87.0	79.5	17.2

scoring), with corresponding standard deviations. Cohen's κ is calculated to assess inter-rater agreement on human evaluation tasks. System metrics are based on 10 randomly selected test queries for each function.

5. Conclusion

This study addresses the scarcity of comprehensive legal knowledge graph data by integrating extensive criminal law theoretical knowledge to construct a Criminal Law Knowledge Graph (CLKG). The CLKG assists non-legal professionals in understanding criminal charges and provides a valuable reference for judicial decision-making. Covering multiple dimensions, including criminal behavior, offenders, and objects of crime, the graph also facilitates case retrieval for legal practitioners. Based on the CLKG, we developed a legal question answering system featuring a user-friendly interactive interface that preserves conversation history and visualizes relevant knowledge subgraphs. The system is adaptable to various downstream applications and provides effective decision support for intelligent legal systems. Future work will focus on incorporating additional criminal charge knowledge, optimizing the interface, expanding the QA system's functionalities, and refining the output quality of legal large language models to enhance overall system performance.

Acknowledgments

The author would like to express sincere gratitude to Professor Yuanyuan Sun from Dalian University of Technology for her guidance and mentorship, which provided a solid foundation for the research presented in this work.

References

Jinze Bai, Shuai Bai, Yunfei Chu, et al. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Baichuan. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023. URL https://arxiv.org/abs/2309.10305.

Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. *arXiv preprint arXiv:1406.3676*, 2014.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023a.

Wanyun Cui, Yanghua Xiao, Haixun Wang, et al. Kbqa: learning question answering over qa corpora and knowledge bases. *arXiv preprint arXiv:1903.02419*, 2019.

Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*, 2023b.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 260–269, 2015.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021.

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, et al. InternLM-law: An open-sourced Chinese legal large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9376–9392, January 2025.

Erwin Filtz. Building and processing a knowledge-graph for legal data. In *The Semantic Web: 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28–June 1, 2017, Proceedings, Part II 14*, pages 184–194. Springer, 2017.

Yuncheng Hua, Yuan-Fang Li, Guilin Qi, Wei Wu, Jingyao Zhang, and Daiqing Qi. Less is more: Data-efficient complex question answering over knowledge bases. *Journal of Web Semantics*, 65:100612, 2020.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*, 2023.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

Feng-Lin Li, Hehong Chen, Guohai Xu, et al. Alimekg: Domain knowledge graph construction and application in e-commerce. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, page 2581–2588, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450368599.

Linfeng Li, Peng Wang, Jun Yan, et al. Real-world data medical knowledge graph: construction and applications. *Artificial intelligence in medicine*, 103:101817, 2020b.

OpenAI. Openai: Introducing chatgpt. 2022. URL https://openai.com/blog/chatgpt.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jeff Z Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu. *Exploiting linked data and knowledge graphs in large organisations*. Springer, 2017.

Hongchen Qin and Yiheng Yao. Agriculture knowledge graph construction and application. In *Journal of Physics: Conference Series*, volume 1756, page 012010. IOP Publishing, 2021.

Yingqi Qu, Jie Liu, Liangyi Kang, Qinfeng Shi, and Dan Ye. Question answering over freebase via attentive rnn with similarity matrix based cnn. *arXiv preprint arXiv:1804.03317*, 38, 2018.

Siva Reddy, Mirella Lapata, and Mark Steedman. Large-scale semantic parsing without question-answer pairs. *Transactions of the Association for Computational Linguistics*, 2: 377–392, 2014.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Nanjing University. Lawgpt. 2023a. URL https://github.com/pengxiao-song/LaWGPT.

Tsinghua University. Lexilaw. 2023b. URL https://github.com/CSHaitao/LexiLaw.

ZY Wang, Qing Yu, Nan Wang, et al. Survey of intelligent question answering research based on knowledge graph. *Computer Engineering and Applications*, 56(23):1–11, 2020.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. Improving question answering over incomplete kbs with knowledge-aware reader. *arXiv preprint arXiv:1905.07098*, 2019.

Xuchen Yao and Benjamin Van Durme. Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 956–966, 2014.

Scott Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP*, 2015.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*, 2023.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Haiguang Zhang, Tongyue Zhang, Faxin Cao, et al. Bca: Bilinear convolutional neural networks and attention networks for legal question answering. *AI Open*, 3:172–181, 2022. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2022.11.002.

Haiguang Zhang, Yuanyuan Sun, Bo Xu, and Hongfei Lin. Legalatle: an active transfer learning framework for legal triple extraction: H. zhang et al. *Applied Intelligence*, 54 (24):12835–12850, 2024.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, et al. Jec-qa: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708, 2020.