electronics

MDPI

*Article*

# SDD-LawLLM: Advancing Intelligent Legal Systems Through Synthetic Data-Driven Fine-Tuning of Large Language Models

**Hanjie Ma [1], Yuhang Lu [1], Zhengdong Xiao [2,*], Jie Feng [1], Haixiang Zhang [1] and Jian Yu [3]**

[1] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; mahanjie@zstu.edu.cn (H.M.); georgeluyh@gmail.com (Y.L.); arlose@zstu.edu.cn (J.F.); zhhx@zstu.edu.cn (H.Z.)
[2] Alibaba Cloud Computing, Hangzhou 310030, China
[3] Hangzhou Codvision Technology, Hangzhou 311100, China; yujian@codvision.com
* Correspondence: zhengdong914@gmail.com

**Abstract:** The extensive use of large language models (LLMs) across various natural language processing tasks has markedly elevated the intelligence of legal systems. Despite their exceptional performance in terms of accuracy, these systems still struggle with explainability. To tackle this challenge, we propose an approach to boost the question-answering abilities of LLMs through data synthesis, focusing on Qwen-7B. By incorporating Retrieval-Augmented Generation (RAG) techniques, we enhance the system's transparency and reliability by introducing detailed reasoning processes (CoT Prompts). Our experimental results indicate that our trained LLMs exhibit significant improvements in both answer accuracy and explainability, especially in objective evaluation tasks. Additionally, subjective assessments reveal that the model's responses are not only precise but also highly understandable, thus boosting user confidence in the system. Overall, our research offers valuable insights and technical advancements for the development of intelligent legal question-answering systems, with significant theoretical and practical implications.

**Keywords:** data synthesis; large language models; intelligent legal systems

check for updates

Academic Editor: Domenico Rosaci

Received: 20 December 2024
Revised: 2 February 2025
Accepted: 12 February 2025
Published: 13 February 2025

**Citation:** Ma, H.; Lu, Y.; Xiao, Z.; Feng, J.; Zhang, H.; Yu, J. SDD-LawLLM: Advancing Intelligent Legal Systems Through Synthetic Data-Driven Fine-Tuning of Large Language Models. *Electronics* **2025**, *14*, 742. https://doi.org/10.3390/electronics14040742



## 1. Introduction

With growing legal awareness and increasing case complexity, ordinary citizens require fast and convenient access to legal advice. Many lack the time or resources for lawyer consultations, particularly in urgent scenarios such as labor disputes, traffic accidents, or contract issues. Prompt professional guidance is highly sought after. Frequent legislative updates at various levels challenge lawyers, causing information overload and raising operational costs.

Despite significant advancements, there remains a critical gap in providing accessible, accurate, and explainable legal assistance to non-experts. The primary objective of this article is to address this gap by developing a model that not only retrieves relevant legal information but also provides clear reasoning and explanations, enhancing user trust and comprehension.

Technological advancements in the legal domain have primarily focused on enhancing Information Retrieval (IR) techniques [1,2], enabling more effective processing of legal texts to address user queries. Systems such as Lexis Answers [3] employ Natural Language Processing (NLP) workflows to deliver answers; however, they struggle with addressing complex questions.

To support the interpretation of statutory terms, Avelka et al. [4] simulated analytical methods historically used for interpreting statutory terms, noting that lawyers often combine manual and computational approaches for this task. They constructed a dataset comprising 42 queries (26,959 sentences), focusing on retrieving short texts (sentences) and specialized document types, such as legal case texts. Experimental results showed that methods ranking sentences directly by their similarity to queries performed poorly. Models such as BM25, TF-ISF, GloVe, and w2vec were evaluated, but transformer-based approaches were not explored.

To understand complex statutory terms, Lima et al. [5] proposed a multi-layered embedding-based retrieval method. They employed the transformer-based text-embedding-3-large model to construct the multi-layered retriever, validating it on the Brazilian Constitution. Results demonstrated that the multi-layered retriever outperformed traditional methods, especially in delivering richer and more precise semantic content.

Despite these advances, there is still a lack of models that integrate reasoning capabilities and explainability tailored specifically to legal queries. Knowledge graph-based systems facilitate the structuring of legal concepts to enable rapid retrieval and reasoning. For example, AILA [6] leverages a legal knowledge graph to enhance question comprehension and answer selection. However, such systems face challenges when handling unstructured and complex questions that demand detailed reasoning.

Machine learning models designed to understand legal contexts, such as BERT [7], RoBERTa [8], and InLegalBERT [9], can address more complex questions by leveraging extensive legal documents for training. However, these models may struggle with multi-entity or uncommon questions and are often hindered by privacy concerns regarding sensitive data.

The emergence of large language models (LLMs), such as ChatGPT, has expanded the capabilities of pre-trained language models, unlocking greater potential for AI applications across diverse domains, including the legal field, economics [10], education [11], and healthcare [12]. Several studies have explored methods to adapt general-purpose large language models (LLMs) for the legal domain, improving their performance on legal tasks. ChatLaw (Cui et al.) [13] introduced a specialized legal model leveraging a Mixture of Experts (MoE) architecture and multi-agent technology, integrating knowledge graphs and human-curated data to enhance accuracy in addressing legal issues. LawGPT (Zhou et al.) [14] incorporated additional pre-training on legal domain-specific data to embed legal knowledge, addressing data privacy concerns and the insufficient legal alignment in open-source models through extensive training on Chinese legal documents and fine-tuning with knowledge-driven instruction datasets. LexiLaw (Li) [15] proposed a fine-tuning approach combining general domain data with specialized legal and document data to refine its understanding of legal language and concepts. Lawyer LLaMA (Huang et al.) [16] constructed an instruction dataset specific to legal tasks and fine-tuned the LLaMA model, leveraging the strengths of LLMs while tailoring them to legal applications. Collectively, these models enhance LLMs' proficiency in understanding and generating legal texts, significantly advancing their capabilities in the legal domain.

For transparency and explainability in intelligent legal systems, we propose SDD-LawLLM, a synthetic CoT data-driven LLM that leverages GPT-4's information extraction to construct a Chain-of-Thought dataset specific to the legal domain. This model is then fine-tuned using this dataset and explores Retrieval-Augmented Generation (RAG) technology to improve explainability.

## 2. Related Work

In this section, we will provide a comprehensive explanation of past approaches aimed at enhancing the explainability of legal AI, delve into Reasoning Enhancement Techniques, analyze their distinctions from previous methods, and ultimately propose our research hypothesis along with the key contributions of this study.

### *2.1. Approaches to Explainable Legal AI Research*

Branting et al. [17] developed the SCALE algorithm, which synergistically integrates structural and semantic patterns from case law corpora to identify textual features that demonstrate both predictive correlations with judicial outcomes and inherent explainability. This methodology effectively automates the extraction of legally salient features from case texts [18], thereby establishing a novel framework for legal decision support systems that achieves explainable predictions without relying on labor-intensive manual feature engineering. While promising in well-structured legal domains, this approach exhibits notable limitations when applied to complex jurisprudence contexts such as criminal law and commercial litigation. These challenging domains necessitate more sophisticated feature extraction pipelines and enhanced modeling architectures to maintain both predictive robustness and high explainability standards. Moreover, the requirement for domain-specific legal expertise in annotating representative case datasets poses practical constraints on the scalability of this method, particularly for resource-limited legal institutions. To address these limitations, we propose a Chain-of-Thought (CoT) data synthesis approach designed to reduce dependency on expert-intensive annotation processes while enhancing explainability in legal reasoning.

Vale et al. [19] examined the extent to which post hoc explainable methods align with the legal definitions of direct and indirect discrimination under European non-discrimination law in post-deployment scenarios. Their study highlights that these methods attempt to approximate the decision-making logic of complex black-box models by constructing simpler surrogate models, thereby facilitating human evaluators' understanding of the models' internal mechanisms. However, the research identifies several significant limitations of these approaches. Firstly, post hoc explainable methods, exemplified by LIME [20], demonstrate inherent instability due to randomness in sampling procedures and methodological variations, leading to inconsistent explanatory outcomes [21]. Secondly, these methods provide only an approximation of the black-box model's behavior, which may fail to accurately capture the true feature space of the original model [22]. Thirdly, they are associated with significant computational overhead [23]. For instance, LIME requires generating numerous samples per instance to construct local models, while Shapley values [24] necessitate the computation of average marginal contributions across all possible feature subsets, resulting in exponentially increasing computational demands as the feature space expands. In response to these limitations, we propose a novel approach that integrates synthesized Chain-of-Thought (CoT) data with Retrieval-Augmented Generation (RAG) techniques. This methodology enhances the model's capability to generate inherently explanatory responses while directly presenting relevant legal provisions. Our approach effectively eliminates the need for post hoc explainability techniques, which typically demand substantial computational resources to reconstruct the reasoning process in a potentially unstable manner.

To improve the reasoning capabilities of large language models (LLMs), Kim et al. [25] developed the Chain-of-Thought (CoT) collection dataset for fine-tuning the T5 model. Their experiments showed a notable enhancement in the model's reasoning performance. Drawing inspiration from their work, we have utilized synthesized CoT data tailored to the legal domain to train LLMs. This approach seeks to enhance both the reasoning capabilities and explainability of models in the legal domain.

### 2.2. Reasoning Enhancement Techniques

Explainability and complex reasoning are crucial for ensuring trustworthiness in legal applications. The Chain-of-Thought (CoT) [26] approach significantly enhances the reasoning capabilities of large language models (LLMs) by generating intermediate reasoning steps, which effectively decompose complex tasks into sequential subtasks. This approach improves performance on multi-step reasoning tasks like legal analysis by refining conditional probability distributions, thereby increasing both the model's explainability and credibility. Retrieval-Augmented Generation (RAG) [27,28] further contributes to this goal by integrating pre-trained language models with retrieval from non-parameterized memory sources, such as dense vector indices of articles, allowing the model to access relevant information during generation. This improves the factual accuracy, specificity, and diversity of generated responses. Additionally, Parameter-Efficient Fine-Tuning (PEFT) [29] techniques, such as LoRA [30], Prefix Tuning [31], Adapter Layers [32], Prompt Tuning [33], and P-tuning [34], have been used. LoRA, in particular, has gained widespread adoption due to its effectiveness and efficiency in fine-tuning LLMs. Together, these methods significantly enhance the explainability, reasoning, and accuracy of models used in legal applications.

### 2.3. Research Variables

#### 2.3.1. Legal Information Retrieval: The Dilemma Between Efficiency and Explainability

Information Retrieval (IR) techniques have evolved to enhance the efficiency of legal text processing and query resolution. Traditional approaches such as BM25 and TF-IDF have demonstrated effectiveness in ranking legal documents [1], yet they often struggle to resolve complex legal queries that require deeper semantic understanding. More recent methods leveraging transformer-based models [5] have enhanced semantic comprehension, yet they still lack the structured knowledge representation required for explainable decision-making.

Knowledge Graph (KG)-based systems, such as AILA [6], have been proposed to address this gap by structuring legal concepts for rapid retrieval and reasoning. However, these systems struggle with unstructured, nuanced legal queries that require deeper contextual analysis. This limitation highlights the need to integrate semantic retrieval with structured knowledge representations, aiming to improve both efficiency and explainability. Therefore, this study adopts Retrieval-Augmented Generation (RAG) [27] technology to simultaneously address both aspects.

#### 2.3.2. Complex Problem-Solving: The Trade-Off Between Structured Reasoning and End-to-End Generation

Addressing complex legal problems necessitates a balance between structured reasoning and generative modeling. Traditional structured reasoning systems, such as AILA [6], rely on explicitly defined legal knowledge, which enhances explainability yet constrains adaptability to novel scenarios. In contrast, LLM-based models like LawGPT [14] exhibit greater adaptability through large-scale training on legal documents; however, they often lack explicit reasoning capabilities, resulting in reduced explainability.

To address this challenge, researchers have explored hybrid approaches. LawGPT incorporates knowledge-driven instruction tuning to enhance domain-specific understanding, yet its reasoning transparency remains limited. An effective solution would integrate structured KG-based reasoning with the generative flexibility of LLMs, ensuring both adaptability and explainability in complex legal reasoning.

#### 2.3.3. Trustworthy AI: Privacy Protection and Explainability Design

Ensuring the trustworthiness of AI in legal applications requires addressing both privacy concerns and model explainability. Large language models (LLMs) are often trained on vast collections of real-world legal documents, raising concerns about the handling of sensitive and confidential data [9]. Privacy-preserving techniques, such as differential privacy and federated learning, have been explored to mitigate data leakage risks; however, these methods often degrade model performance.

Regarding explainability, techniques such as Chain-of-Thought (CoT) reasoning [26] have demonstrated improved transparency by explicitly generating intermediate reasoning steps. Similarly, Retrieval-Augmented Generation (RAG) [27] enhances factual accuracy by incorporating external knowledge sources during generation. However, existing implementations fail to adequately balance privacy-preserving data handling with explainable reasoning, underscoring the need for a more comprehensive approach.

#### 2.3.4. Summary of Research Gaps

Current legal AI systems face challenges in balancing three critical factors: retrieval efficiency, explainability, and complex problem-solving. While Information Retrieval (IR) techniques optimize efficiency, they lack depth in legal reasoning. Large language models (LLMs) improve semantic understanding but often struggle to generate structured and interpretable justifications. Moreover, addressing complex legal problems requires not only advanced reasoning capabilities but also the integration of multiple knowledge sources to ensure consistency and reliability in decision-making.

To bridge these gaps, we propose integrating Retrieval-Augmented Generation (RAG) with synthetically generated Chain-of-Thought (CoT) data. This hybrid approach seeks to enhance explainability while improving the model's ability to tackle intricate legal challenges, ultimately increasing the trustworthiness and efficacy of legal AI systems. Future research will focus on enhancing the reliability and adaptability of intelligent legal systems in handling multifaceted legal scenarios.

### *2.4. Research Hypotheses*

Based on the literature analysis, this study proposes the following hypotheses:

**H1:** *Synthetic CoT data, by explicitly modeling the legal reasoning chain, can significantly improve the answer accuracy and explainability of LLMs.*

**H2:** *RAG technology, by dynamically retrieving from legal knowledge bases, can compensate for the limitations of pure parameter fine-tuning in factual answering, but may reduce answer relevance due to retrieval noise.*

**H3:** *Parameter-efficient fine-tuning (LoRA) can effectively adapt to the legal domain while preserving the general capabilities of the model, preventing catastrophic forgetting.*

Our contributions include the following:

a. Proposing a synthesis data-driven fine-tuning method for LLMs applied to Chinese intelligent legal systems.
b. Creating a legal Chain-of-Thought dataset and utilizing efficient fine-tuning techniques to boost LLM performance.
c. Demonstrating superior performance on major legal tasks through experimental results, providing strong evidence for our method's effectiveness.

## 3. Methodology

### *3.1. Dataset*

Our method aims to comprehensively enhance the domain reasoning capabilities of large language models (LLMs) to improve both the answer quality and explainability of intelligent legal systems. Thus, we constructed a high-quality legal CoT dataset and combined it with a general-purpose dataset (COIG-CQIA [35]) for training SDD-LawLLM. The first part aims to incorporate legal reasoning capabilities into the LLM, while the second part helps maintain the model's original abilities without degradation after training. The selection of these datasets is based on their relevance and prior validation in similar tasks [25,36]. The process of constructing the dataset is illustrated in Figure 1.

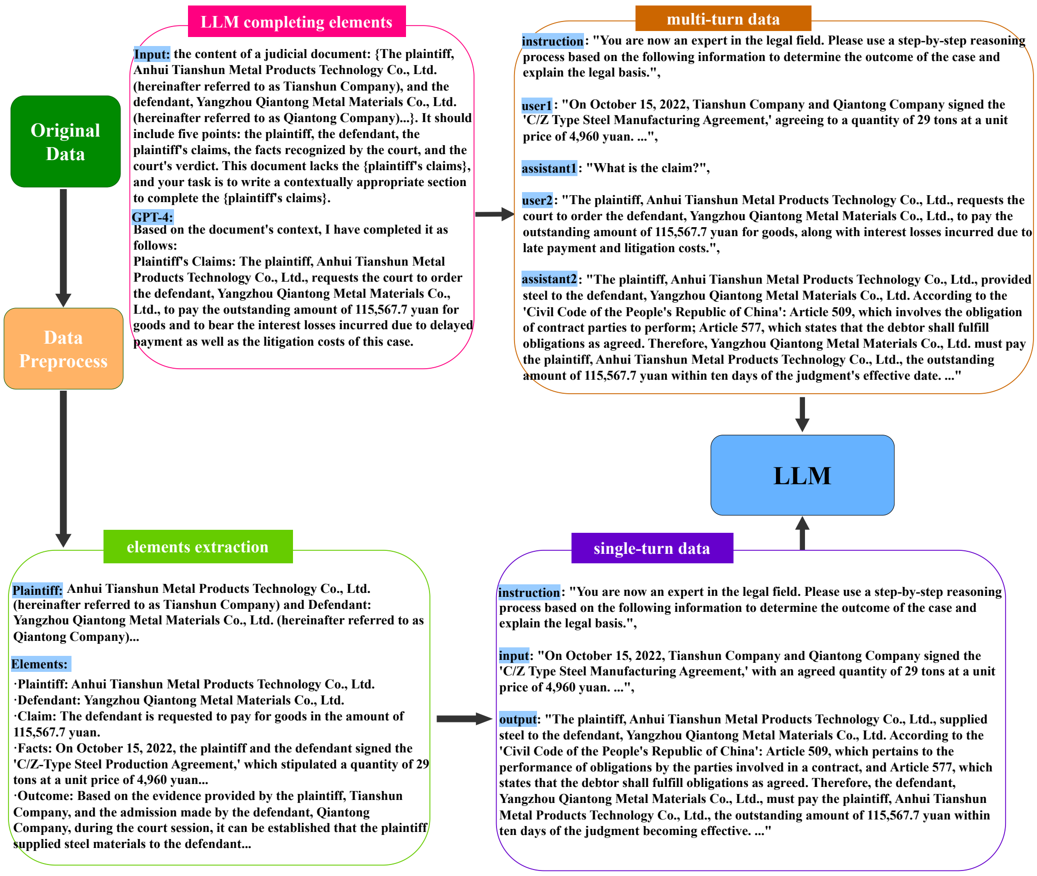


**Figure 1.** The process of data synthesis.

#### 3.1.1. Data Preprocessing

In our study, the data for the self-constructed legal domain question-answering dataset were sourced from China Judgements Online (CJO) (https://wenshu.court.gov.cn/, accessed on 17 April 2024), focusing on publicly available judicial documents related to civil and compensation cases. During data preprocessing, we excluded documents with the following characteristics: (1) Short Content, where documents were too brief due to lawsuit withdrawals or settlements, as they lacked the necessary logical structure for constructing Chain-of-Thought (CoT) data; and (2) Ambiguous References, such as retrial judgments that repeatedly cite initial trial content, resulting in unclear references. An example of this is phrases like "legally overturning the judgment to support all the appellant Supply and Marketing Company's claims from the first instance" which contain ambiguous references that would be unclear to the model. For short content, we developed a Python script to filter out data with content less than one page. For ambiguous references, we used regular matching to filter out data where titles contain keywords such as "second trial", "third trial", and "retrial". This approach ensures a cleaner, more effective dataset for analysis.

#### 3.1.2. Data Synthesis

Using the preprocessed data, we constructed Chain-of-Thought (CoT) data tailored for the legal domain. Given that legal question-answering systems operate in serious contexts, it is crucial that the model's responses to user queries are explainable. This means clearly articulating the reasoning behind predictions in an understandable way, establishing trust and ensuring users can grasp the rationale without needing specialized legal knowledge.

User queries can be categorized into two main types: knowledge-based questions and reasoning-based questions, as shown in Figure 2. Knowledge-based questions are effectively supported by a large knowledge base. On the other hand, reasoning-based questions benefit from the logical reasoning capabilities provided by Chain-of-Thought (CoT) technology. These questions typically include two key characteristics: (1) a statement of relevant facts, and (2) an inquiry about the validity of a claim, where the user asks whether their claim or request can be satisfied. This categorization helps tailor the system's approach to answering, ensuring both accuracy and comprehensibility.

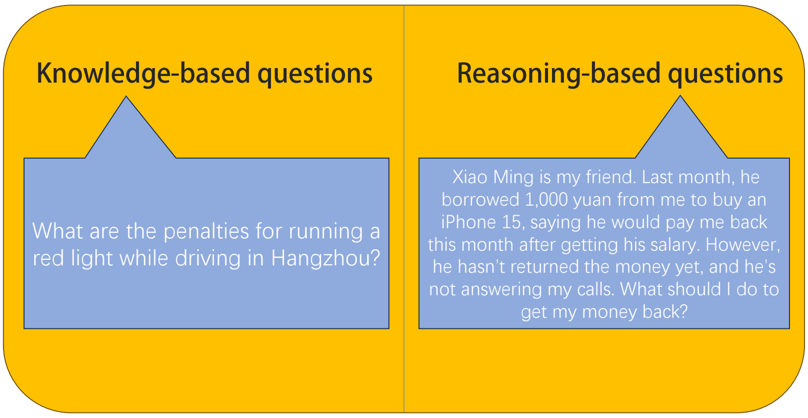


**Figure 2.** Knowledge-based questions and reasoning-based questions.

To address the user's need for determining the validity of their claim based on presented facts, we propose a structured method for constructing Chain-of-Thought (CoT) data that facilitates logical reasoning grounded in legal principles.

Data Extraction

A judicial document typically consists of five key elements: plaintiff, defendant, plaintiff's claim, court-recognized facts, and the court's decision. Constructing CoT data involves identifying the logical relationships between these elements and utilizing them to build the dataset.

Therefore, we developed a program to extract essential elements from judicial documents—plaintiff, defendant, plaintiff's claim, court-recognized facts, and court's decision. Direct extraction can be performed for parties involved and the claim, while court-recognized facts and decisions require pattern matching based on common phrasing such as "After review by this court" or "As determined by this court".

Due to the varying quality of judicial documents and imperfections in the extraction process, some extracted results may lack elements. We prompted GPT-4 to complete missing elements in incomplete entries. Specific prompts guide the AI in generating plausible content for the five key elements, as shown in Table A1.

Data Construction

For each dataset entry, we establish a clear "instruction" to guide the model's reasoning process. This instruction serves as a command memory for legal Q&A, addressing the issue of "catastrophic forgetting" during fine-tuning and aiding the model in recalling relevant information. The instruction we use is as follows: "You are now an expert in the legal field. Please base your judgment on the following information, using a step-by-step reasoning process to determine the outcome of the case and explain the legal basis".

To address situations where user inputs are missing elements, we use incomplete data that have undergone element completion to construct Multi-Round Dialogues. To construct multi-round dialogues for incomplete data, Step 1 involves using the available court-recognized facts or the plaintiff's claim as the initial input ("user1"). In Step 2, the system prompts for missing elements ("assistant1")—e.g., "What is your claim?". Step 3 requires the user to supplement the missing information ("user2"). Finally, in Step 4, key facts, legal provisions, and judgment outcomes are extracted from the court's decision using GPT-4-based prompts, as shown in Table A2, and formatted as responses with the following structure: "In this case, fact, based on legal provision, therefore judgment outcome" ("assistant2").

For complete data, we integrate court-recognized facts and the plaintiff's claim into a single "input". Extract and format key information from the court's decision similarly to Step 4, serving as the "output".

The final data entry formats are streamlined as follows, and you can see the templates in Figures A1 and A2:

For incomplete data: {"instruction":. . . , "user1":. . . , "assistant1":. . . , "user2":. . . , "assistant2"; . . . }.

For complete data: {"instruction":. . . , "input":. . . , "output": . . . }.

### 3.2. SDD-LawLLM

To build an intelligent legal system with robust reasoning and retrieval capabilities, we developed SDD-LawLLM using parameter-efficient fine-tuning and a Retrieval-Augmented Generation (RAG) approach. The selection of Qwen 7B as the base model is justified by its superior performance in benchmark tests [37] compared to other models, and because newer models may have been trained on evaluation datasets during their pretraining phase, potentially skewing results. The workflow of the legal question-answering system, as illustrated in Figure 3, consists of three key stages. First, in the **Knowledge Base Construction** stage, raw legal documents are split and organized into a structured legal domain knowledge base, providing the system with access to a comprehensive repository of legal information. Next, in the **CoT Dataset Development** stage, we construct a Chain-of-Thought (CoT) dataset specifically designed for training the large language model (LLM). This dataset enhances the model's ability to perform logical reasoning within the legal context, ensuring accurate and explainable responses. Finally, the **Retrieval-Augmented Generation (RAG)** stage involves utilizing advanced retrieval algorithms to dynamically access additional legal domain knowledge, augmenting the LLM's answer generation process and ensuring responses are enriched by up-to-date and relevant legal information.

#### 3.2.1. Training

We chose Qwen 7B as the base model primarily due to its widely recognized performance, and secondly because newer models may have been trained on evaluation datasets during their pretraining phase. We employ Parameter-Efficient Fine-Tuning (PEFT) techniques to capture domain-specific patterns and characteristics in the legal field, thereby enhancing the question-answering capabilities and explainability of the model. For this work, we adopt Low-Rank Adaptation (LoRA), one of several PEFT optimization methods available, including Prefix Tuning and Adapter Layers.

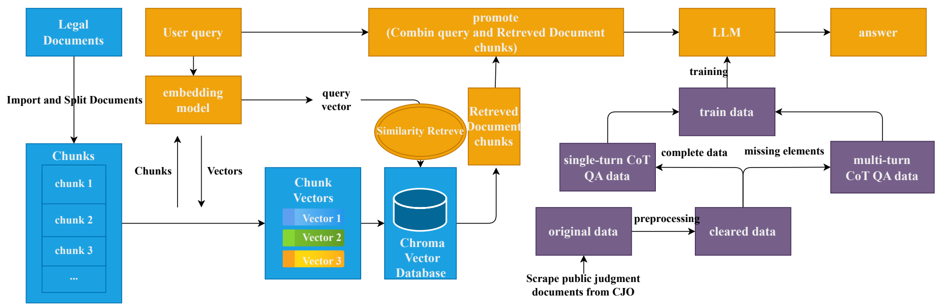


**Figure 3.** Legal question-answering system workflow. The blue module represents the construction of the retrieval database, the purple module represents LLM training, and the yellow module represents the inference module.

LoRA optimizes the fine-tuning process by decomposing the weight matrix $\Delta W$ into two low-rank matrices $A$ and $B$. This decomposition preserves the model's performance while significantly reducing the number of parameters to be fine-tuned. We selected LoRA based on its proven efficiency and effectiveness in similar tasks [15]. During training, the original model parameters are frozen, and the $A$ matrix is initialized using Kaiming initialization, while the $B$ matrix is initialized to zero. This approach ensures that the model can efficiently learn the specific nuances of the legal domain without overfitting or losing the general knowledge captured by the pre-trained model.

$$W^{'} = W + \Delta W \tag{1}$$

$$\Delta W = A \times B \tag{2}$$

$$W^{'} = W + A \times B \tag{3}$$

Here, $W^{'}$ is the updated weight matrix, $W$ is the original weight matrix, and $\Delta W$ is a weight matrix of the same size as $W$.

We trained our models using eight NVIDIA 4090 GPUs, based on the Qwen 7B. For supervised fine-tuning in the legal domain, we mixed our self-constructed 2 K legal CoT data with 20 K general data (in a 1:10 ratio) and used LoRA technology to fine-tune Qwen 7B. The LoRA rank was set to 8, alpha to 16, dropout to 0.05, learning rate to 0.0005, batch size to 16, and the number of training epochs to 30.

#### 3.2.2. RAG

The Retrieval-Augmented Generation (RAG) process in our system is meticulously designed to enhance the accuracy and relevance of legal question-answering through four key steps. In the first step, legal documents, primarily in Word format, are segmented into manageable chunks using Python scripts and the DOCX framework, with each chunk consisting of 500 characters and a 20% overlap to maintain contextual integrity. Tables and formatted data are converted into JSON, summarized into single sentences, and the original table positions are replaced with these summaries, improving embedding quality and retrieval accuracy. In the second step, the document chunks are converted into vectors using the BGE-M3 embedding model, selected for its superior performance in multilingual embedding benchmarks [38], and stored in Chroma, an AI-native open-source vector database. Chroma's unique feature of separating vector storage from the original documents while maintaining the same UUID ensures retrieval results return the original content, preserving information completeness. In the third step, user queries are embedded, and relevant document chunks are retrieved based on text similarity, using cosine similarity, which outperforms other metrics in context-specific evaluations [39]. Retrieved results are reranked to select the top-K most relevant documents, with K = 5 based on F1-score optimization, ensuring a balance between relevance and response speed. In the final step, the selected relevant chunks and user query are combined into a structured prompt, which is then fed into the LLM to generate a final, reasoned response. This approach ensures the system generates highly accurate and contextually relevant answers to legal questions.

By integrating these steps, our RAG process not only leverages the strengths of advanced embedding models and efficient databases but also ensures that the generated responses are grounded in accurate, contextually rich legal information. This method significantly enhances the reliability and explanatory power of the legal question-answering system.

## 4. Evaluation

In August 2023, several research institutions and universities released a joint proposal [40] suggesting the establishment of a comprehensive evaluation system that combines both objective and subjective metrics. Objective metrics are assessed through the model's performance in eight legal knowledge applications [37], while subjective metrics are evaluated by legal experts. This approach ensures a more robust and comprehensive evaluation of the model.

### *4.1. Subjective Evaluation*

As shown in Table 1, the model's responses were evaluated based on correctness, explainability, and relevance scores provided by legal experts. The scoring method is outlined in Equation (4). For this evaluation, we invited two legal experts to manually score 200 responses for each method across these three dimensions.

As detailed in Table 2, SDD-LawLLM significantly outperforms the base LLM in accuracy, explainability, and relevance. This validates H1, emphasizing that synthetic CoT data can substantially improve the answer accuracy and explainability of large language models, in line with the findings of Kim et al. [25]. When utilizing RAG, both accuracy and explainability exhibited improvements; however, a slight decline in relevance was observed. This finding supports hypothesis H2 and can be attributed to RAG incorporating additional information that, while relevant to the query, may not always be directly pertinent to addressing the specific question. Interestingly, Qwen 7B with RAG performed better than SDD-LawLLM in certain aspects. However, it is important to note that Qwen 7B is a general-domain model, which inherently lacks strength in answering domain-specific questions such as those in the legal field. RAG technology compensates for this shortcoming by providing the LLM with relevant knowledge to answer questions more effectively. In contrast, the primary goal of training SDD-LawLLM is to enhance its complex reasoning ability within the legal domain, rather than improving its capability to answer domain-specific questions alone.

$$Score = (S_1 + S_2 + S_3)/5 \tag{4}$$

Here, $S_1$: correctness score; $S_2$: explainability score; $S_3$: relevance score.

### *4.2. Objective Evaluation*

In this section, we evaluate SDD-LawLLM's performance across seven legal applications based on the LawBench benchmark [37]. These applications include three knowledge-based tasks (#1–#3) and four reasoning-based tasks (#4–#7), specifically the following: fact-based article prediction (#1), scene-based article predic-tion (#2), criminal damages calculation (#3), marital disputes identification (#4), charge prediction (#5), case analysis (#6), and consultation (#7). Under a zero-shot setting, we compare SDD-LawLLM against lawGPT, lawyer LLaMA, ChatLaw 13B, LexiLaw, and Qwen 7B. The comparative results are summarized in Table 3.

**Table 1.** Subjective evaluation criteria.

| Score | Correctness Score | Explainability Score | Relevance Score |
|---|---|---|---|
| 5 | The response error is within 5%, and the judgment result is completely accurate. | The decision-making and reasoning process is clear and entirely convincing | The decision reasoning results can reflect all useful correlations in the conditional information provided by the user. |
| 4 | The error exceeds 5% but does not exceed 10%; the result is approximately correct. | The decision-making and reasoning process is over 80% clear and convincing. | The decision-making and reasoning results reflect 80% or more of the relevance of the condition information provided by the user. |
| 3 | The error exceeds 10% but does not exceed 15%; the result is basically correct. | The decision-making and reasoning process is clear and convincing in more than 60% but less than 80% of cases. | The decision-making and reasoning results reflect 60% or more but less than 80% of the relevance of the condition information provided by the user. |
| 2 | The error exceeds 15% but does not exceed 20%; the result is partially correct. | The decision-making and reasoning process is clear and convincing in more than 40% but less than 60% of cases. | The decision-making and reasoning results reflect 40% or more but less than 60% of the relevance of the condition information provided by the user. |
| 1 | The error exceeds 20% but does not exceed 25%; the result is partially correct. | The decision-making and reasoning process is clear and convincing in more than 20% but less than 40% of cases. | The decision-making and reasoning results reflect 20% or more but less than 40% of the relevance of the condition information provided by the user. |
| 0 | The error exceeds 25%; the result is incorrect. | The decision-making and reasoning process is entirely unclear and difficult to be convincing. | The decision-making and reasoning results do not reflect the relevance of the condition information provided by the user. |

**Table 2.** Results of the subjective evaluation.

| Methods | Correctness | Explainability | Relevance | Score |
|---|---|---|---|---|
| Qwen 7B | 2.24 | 3.40 | 4.12 | 1.952 |
| SDD-LawLLM | 2.98 | 3.84 | 4.22 | 2.208 |
| Qwen 7B with RAG | 3.46 | 4.20 | 3.64 | 2.26 |
| SDD-LawLLM with RAG | 3.94 | 4.64 | 3.98 | 2.51 |

**Table 3.** A comparison of the performance of various models in a zero-shot setting. The models include SDD-LawLLM, lawGPT, lawyer LLaMA, chatlaw 13B, LexiLaw, and Qwen 7b. Qwen 7b is a general-purpose model, while the rest are legal domain models. The model with the best performance in each task is highlighted in bold.

| Models | Task 1 | Task 2 | Task 3 | Task 4 | Task 5 | Task 6 | Task 7 | avg | newTask 2 |
|---|---|---|---|---|---|---|---|---|---|
| lawGPT | 0.15 | 11.01 | 15.4 | 15.68 | 6.2 | 6.85 | 7.62 | 8.01 | 0 |
| lawyer LLaMA | 0.6 | 25.94 | 39.2 | 31.3 | 17.8 | 15.88 | **16.94** | 12.83 | 0 |
| chatlaw 13B | 5.35 | 26.02 | **41.6** | 12.73 | **34.2** | **42.24** | 16.55 | 19.49 | 0 |
| LexiLaw | 13.15 | **35.78** | 35.8 | 39.99 | 20.8 | 15.6 | 15.82 | 25.28 | 0.012 |
| Qwen 7b | 19.88 | 20.8 | 32.8 | 20.16 | 0.4 | 7.81 | 10.07 | 15.99 | 0.006 |
| SDD-LawLLM | **23.12** | 11.61 | 31.2 | **49.86** | 24.8 | 35.6 | 12.33 | **26.93** | **0.024** |

The results indicate that SDD-LawLLM outperforms the base model in primary tasks and achieves the best average performance among legal domain models, despite having a slight parameter disadvantage compared to ChatLaw 13B on a few tasks.

Notably, Qwen 7B, when optimized with our method, demonstrates significant improvements across multiple tasks, especially in reasoning-based tasks (#4–#7), where its logical analysis capabilities are bolstered by Chain-of-Thought (CoT) data integration.For knowledge-based tasks (Task 1 and Task 3), performance aligns with the base model. This finding supports hypothesis H3.

However, there is a noted decline in scene-based article prediction (#2). Analysis revealed that SDD-LawLLM generates longer responses than the relatively brief target texts, leading to lower ROUGE-L scores due to length mismatch. To address this, we re-evaluated Task 2 using regular expression matching to check for correct legal article outputs. This adjustment is reflected in the new Task 2 column in Table 3, showing improved performance for SDD-LawLLM. Most models scored 0 on this task because of the strict requirement to accurately output both the legal code name and specific article number.

### 4.3. Evaluation Summarize

Through subjective and objective evaluations, we confirmed that our method enables LLMs to acquire specific domain knowledge and reasoning skills, thereby enhancing response accuracy and explainability across various base model architectures. Our approach maintains the model's inherent legal domain capabilities while competing effectively with other legal models.

In summary, our method significantly boosts the accuracy and transparency of legal question-answering, aligning with the key objectives of intelligent legal systems.

## 5. Conclusions

In this study, we fine-tuned large language models (LLMs) using synthetic CoT data and implemented the Retrieval-Augmented Generation (RAG) method to develop an intelligent legal system. Our results confirm that this approach improves response accuracy, preserves the foundational strengths of the models, and enhances explainability by incorporating detailed reasoning processes. These findings highlight the method's ability to address the need for transparency and reliability in legal applications.

### 5.1. Practical Implications

The developed intelligent legal system demonstrates significant potential for real-world applications. By improving response accuracy and explainability, legal professionals can leverage this technology to streamline case analysis, draft legal documents, and provide more informed legal opinions. Additionally, the integration of detailed reasoning processes facilitates better understanding and trust among users, promoting broader adoption of AI-driven solutions in the legal domain. The system's ability to handle standard legal queries effectively also indicates its utility in enhancing accessibility to legal information for the general public.

### 5.2. Theoretical Implications

From a theoretical standpoint, this research contributes to the growing body of knowledge on the integration of synthetic CoT data and RAG methods in enhancing LLMs. The findings underscore the importance of detailed reasoning processes in improving model explainability and reliability, offering a foundation for future studies aiming to refine AI-driven systems in complex domains. Moreover, this work illustrates the interplay between model fine-tuning and retrieval-augmented techniques, providing insights into optimizing LLM performance for specialized applications.

### 5.3. Limitations

While the proposed system performs well in standard scenarios, its effectiveness in handling complex legal contexts remains limited, especially in cases involving ambiguous or multi-layered regulations. The variability of legal frameworks and languages across jurisdictions presents significant challenges for broader implementation, requiring extensive domain-specific customization. Additionally, deploying such systems in real-world applications necessitates careful consideration of ethical implications and compliance with local regulations, particularly regarding privacy, accountability, and fairness.

### 5.4. Future Work

To address these challenges, future research will focus on several key areas:

1. Enhancing Robustness in Complex Scenarios: Improving the system's ability to navigate intricate legal contexts through advanced reasoning and contextual understanding.
2. Developing Adaptable Frameworks: Creating flexible architectures that can efficiently incorporate diverse legal systems and multilingual capabilities.
3. Exploring Ethical AI Frameworks: Ensuring transparency, accountability, and compliance with legal norms by integrating ethical guidelines into the deployment process.

By pursuing these areas, we aim to further bridge the gap between advanced AI systems and the nuanced demands of real-world legal practice, fostering the development of more reliable and universally applicable legal AI solutions.

**Author Contributions:** Conceptualization, H.M.; methodology, Y.L. and H.M.; software, Y.L.; validation, Y.L. and Z.X.; formal analysis, Z.X.; investigation, H.M.; resources, H.M. and H.Z.; data curation, J.F.; writing—original draft preparation, Y.L.; writing—review and editing, Z.X. and H.M.; visualization, J.F. and J.Y.; supervision, H.M. and J.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Hangzhou Science and Technology Bureau under grant number 2022AIZD0145. The APC was funded by Hanjie Ma.

**Data Availability Statement:** Due to privacy, data sharing is not applicable to this article.

**Conflicts of Interest:** Author Zhengdong Xiao is employed by the company Hangzhou Alibaba Cloud Computing Technology Co., Ltd. and author Jian Yu is employed by the company Hangzhou Codvision Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

## Appendix A

### Appendix A.1

The prompts used in the CoT data synthesis process.

**Table A1.** Prompt to complete the missing elements.

| |
|---|
| Prompt: This is a segment from a judicial document: {original document}, which should include five key elements: 1. the plaintiff, 2. the defendant, 3. the plaintiff's claims, 4. the facts determined by the court, 5. the court's judgment.<br>This document is missing {missing part}. Your task is to write a contextually appropriate section to complete the {missing part}. |

**Table A2.** An example extracted by GPT-4.

| |
|---|
| Prompt:<br>"The legal loan relationship is protected by law, and the debt should be repaid. In this case, the defendant Hu and his wife settled their previous debts with the plaintiff and signed a new "Personal Loan Contract" with the plaintiff Bao, and then issued a "Receipt of Loan" to the plaintiff, which is the defendant's true intention. Although the defendant claimed that the initial loan was a total principal of 13,000 yuan, and the "Personal Loan Contract" signed on 13 November 2019 was issued after compound interest calculation, in order to provide any relevant evidence to support it, this court supports the plaintiff's request for the defendant Hu to repay the loan. This court does not accept the defendant's defense. As for the interest part, the plaintiff's claim does not exceed the interest rate. The upper limit stipulated by laws and regulations is therefore supported by this court. In summary, based on Article 1 (2) of the "Several Provisions of the Supreme People's Court on the Application of Time Effect", Articles 60, 196, 206, and 207 of the "Contract Law of the People's Republic of China", Article 25, Article 28 (2) (2), Article 31 of the "Provisions of the Supreme People's Court on Several Issues Concerning the Application of Law in the Trial of Civil Lending Cases" (revised on 23 December 2020), and Article 67 (1) of the "Civil Procedure Law of the People's Republic of China", the judgment is as follows: the defendant Hu shall immediately pay the plaintiff Bao the loan principal of 49,900 yuan and interest of 24,755.39 yuan [49,900 yuan $\times$ 1.21% $\times$ 41] after this judgment takes effect. Month (from 13 December 2019 to 12 May 2023) The total amount is 74,655.39 yuan and the interest will continue to be calculated at a monthly interest rate of 1.21% based on the loan principal of 49,900 yuan from 13 May 2023 until full repayment".<br>The above text contains multiple facts, legal provisions, and judgment results. Please help me sort them out and correspond them. |
| GPT-4:<br>Correspondence between facts and legal provisions Loan relationship: There is a legal loan relationship between the defendant Mr. and Mrs. Hu and the plaintiff Mr. Bao, and after settlement, they signed a Personal Loan Contract.<br>Legal provisions: According to Article 60 of the Contract Law of the People's Republic of China, the performance of a contract shall follow the principle of good faith.<br>Loan amount: The initial loan principal was 13,000 yuan, but the subsequent contract included 57,900 yuan, which may have been calculated as compound interest.<br>Legal terms: According to the relevant provisions of the Contract Law of the People's Republic of China (such as Article 196), the terms of the loan contract are stipulated.<br>Lack of evidence: The defendant failed to provide evidence to support their defense.<br>Legal provisions: According to Article 31 of the "Provisions of the Supreme People's Court on Several Issues Concerning the Application of Law in the Trial of Civil Lending Cases", the defendant shall provide evidence to support its defense.<br>Interest portion: The plaintiff's claim that the interest rate does not exceed the legal limit is supported by the court.<br>Legal provisions: According to Article 25 and Article 28 (2) (ii) of the "Provisions of the Supreme People's Court on Several Issues Concerning the Application of Law in the Trial of Civil Lending Cases", the interest shall not exceed the upper limit prescribed by law. The judgment result corresponds to the legal provisions The defendant shall immediately pay the loan principal of 49,900 yuan and interest of 24,755.39 yuan, totaling 74,655.39 yuan<br>Legal provisions: In accordance with Article 67 (1) of the Civil Procedure Law of the People's Republic of China and relevant provisions of the Contract Law, support the plaintiff's request. Continue to calculate interest at a monthly rate of 1.21% until full repayment<br>Legal provisions: According to Article 28 of the "Provisions of the Supreme People's Court on Several Issues Concerning the Application of Law in the Trial of Civil Lending Cases", the legality of interest calculation is supported. |

*Appendix A.2*

Multi-Round Data Template and Single-Round Data Template

**Multi-Round Data Template**

"**instruction**":" You are now a legal expert. Based on the following information, please use a step-by-step reasoning process to determine the outcome of the case and explain the legal basis."
"**user1**":" On January 15, 2021, the Industrial and Commercial Bank of China, Anda Branch, and Heilongjiang Anruijia Transportation Co., Ltd. signed a *Working Capital Loan Agreement*, which stipulated a loan of 30,000,000 yuan for a term of 12 months with an annual interest rate of 4.35%. The contract also specified that overdue interest would be compounded. On January 19, 2021, the bank disbursed the loan of 30,000,000 yuan to Heilongjiang Anruijia Transportation Co., Ltd.
As of now, the loan remains unpaid. Heilongjiang Anruijia Transportation Co., Ltd. owes 398,750 yuan in interest, with compounded overdue interest of 1,309.24 yuan, bringing the total amount due to 30,400,059.24 yuan (principal and interest)."
"**assistant1**" :" What is your claim?"
"**user2**" :" The defendant, Heilongjiang Anruijia Transportation Co., Ltd., shall immediately repay the loan principal of RMB 30,000,000; the defendant shall pay interest of RMB 398,750 up to the date of filing the lawsuit; the defendant shall pay overdue interest compounded at RMB 1,309.24; the defendant shall pay interest and overdue interest compounded from the date of the lawsuit until the actual repayment date (calculated based on the contract's agreed interest rate); the defendant shall bear all litigation costs of this case."
"**assistant2**" :"In this case, Heilongjiang Anruijia Transportation Co., Ltd., as the borrower, signed a "Loan Agreement" with the Andai Branch of Industrial and Commercial Bank of China, with a loan amount of RMB 30,000,000. According to Article 677 of the Civil Code of the People's Republic of China, the contract is legally established and valid, and all parties should perform their obligations as agreed. Therefore, the court rules that Heilongjiang Anruijia Transportation Co., Ltd. should repay the loan principal of RMB 30,000,000 along with the corresponding interest to the Andai Branch of Industrial and Commercial Bank of China.
In this case, Zhao Wenlong and Kong Fanyun signed a "Guarantee Agreement," providing joint and several guarantees for a debt of RMB 25,000,000; Heilongjiang Anruijia New Materials Technology Development Co., Ltd. signed a "Maximum Guarantee Agreement," providing joint and several guarantees for a debt of RMB 33,000,000; Zhao Wenlong and Kong Fanyun also signed another "Maximum Guarantee Agreement," providing joint and several guarantees for a debt of RMB 60,000,000. According to Articles 16 and 21 of the Guarantee Law of the People's Republic of China, if the debtor fails to fulfill repayment obligations, it constitutes a breach of contract, and the guarantors should bear joint and several liability within the scope of the guarantee. Therefore, the court rules that Zhao Wenlong, Kong Fanyun, and Heilongjiang Anruijia New Materials Technology Development Co., Ltd. should bear joint and several liability for the above debts within the scope agreed upon in their respective "Guarantee Agreements" and "Maximum Guarantee Agreements"."

**Figure A1.** Multi-round data template.

**Single-Round Data Template**

"**instruction**": " You are now a legal expert. Based on the following information, please use a step-by-step reasoning process to make a judgment on the case outcome and explain the legal basis."

"**input**": " The defendant, Aixin Education Third Branch, is a subsidiary of the defendant, Aixin Education Company. On September 4, 2021, the plaintiff's father, Yu Zhitao, signed a *Course Registration Agreement* with the defendant Aixin Education Third Branch on behalf of the plaintiff. It was agreed that the plaintiff would take ballet training courses provided by the defendant, totaling 144 class hours, at the Tianjin Ao City location. The total amount payable was 24,350 yuan, with an actual payment of 20,950 yuan. After signing the agreement, Yu Zhitao paid the course fee of 20,950 yuan to the defendant via WeChat.
The plaintiff stated that due to the relocation of the campus from Ao City in Nankai District to Caiyue City in Hexi District, Aixin Education Third Branch offered two options: a discounted continuation of the course or a course cancellation with a refund. Since the plaintiff did not use the course package and the new campus at Caiyue City was farther away, the plaintiff filled out a membership refund approval form on December 28, 2022, as requested by Aixin Education Third Branch. The form stated that the refund was due to the relocation of the campus and for unused course packages, with a refund amount of 20,950 yuan. However, the defendant Aixin Education Third Branch has not refunded the amount to the plaintiff. Plaintiff's Claims: 1.Request the court to terminate the *Course Registration Agreement* between the plaintiff and Aixin Education Third Branch. 2.Request the court to order Aixin Education Third Branch to refund the remaining course fee of 20,950 yuan, with Aixin Education Company bearing joint and several liability. 3.Request the defendants to bear all litigation costs and pay them directly to the plaintiff."

"**output**" : " In this case, the plaintiff and the defendant, Aixin Education Third Branch, signed a dance training service contract in September 2021, with the plaintiff paying a total training fee of 20,950 yuan. On October 28, 2022, due to the defendant's campus relocating, the plaintiff opted to terminate the contract and requested a refund. The campus manager and headquarters representative of the defendant promised that the contract would be terminated due to the relocation, and the full amount of the training fee, 20,950 yuan, would be refunded to the plaintiff. However, after the plaintiff signed the refund notice, the defendant failed to fulfill their commitment to return the funds. According to Articles 467(1), 562(1), and 586(1) of the *Civil Code of the People's Republic of China*, the defendant's campus relocation rendered the contract unfulfillable as agreed, and the parties failed to reach an agreement on changing the service location. The plaintiff is therefore entitled to terminate the contract and demand a refund of the unused portion of the training fees. As a result, the *Course Registration Agreement* signed between the plaintiff and Aixin Education Third Branch is hereby terminated. The defendant, Aixin Education Third Branch, is ordered to refund the unused training fees of 20,950 yuan to the plaintiff. "

**Figure A2.** Single-round data template.

*Appendix A.3*

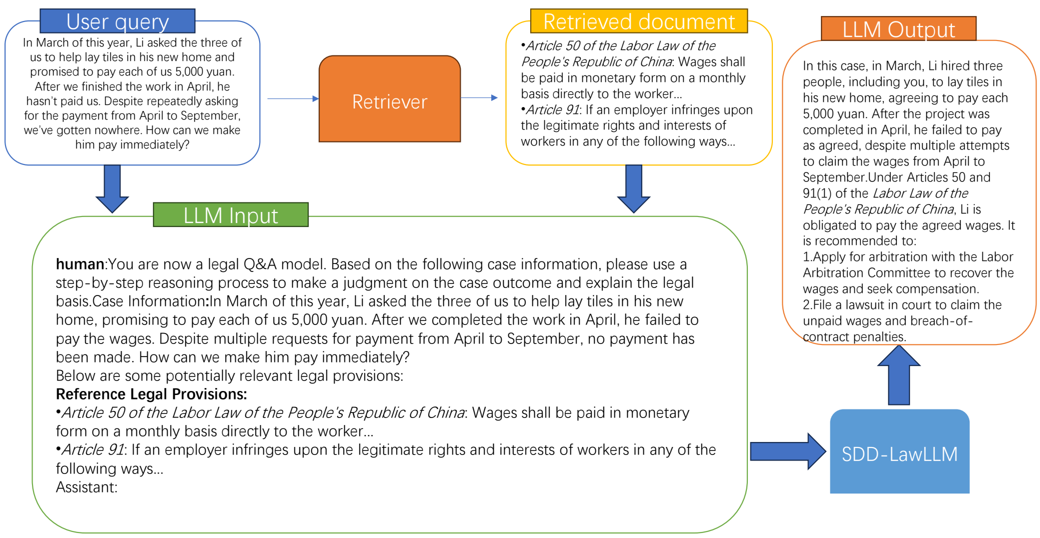


**Figure A3.** The inference process of SDD-LawLLM.

*Appendix A.4*

# The People's Court of Lingqiu County, Shanxi Province
# Civil Judgment

(2023) Jin 0224 Minchu No. 241

Plaintiff: Bai, residing in Chaoyang District, Beijing. Authorized Litigation Agent: Ge, lawyer at Beijing Lianggao Law Firm.

Defendant: Zhang, residing in Lingqiu County. Authorized Litigation Agent: Liu, lawyer at Shanxi Xiangguang Law Firm.

Defendant: Wang, residing in Sanhe City, Hebei Province.

The plaintiff, Bai, filed a lawsuit against the defendants, Zhang and Wang, concerning a dispute over a donation contract. The court accepted the case on April 14, 2023, and applied the simplified procedure, holding a public hearing. The plaintiff's authorized litigation representative, Ge, as well as defendants Zhang, Wang, and their respective representatives, Liu, attended the court proceedings. The case has now been concluded.

The plaintiff, Bai, requests the court to: 1. Confirm that the donation of 338,676.12 RMB from defendant Wang to defendant Zhang between September 16, 2022, and March 20, 2023, is invalid; 2. Order defendant Zhang to return the donation amount of 338,676.12 RMB to the plaintiff; 3. Require the defendants to pay the litigation costs of this case. The facts and reasons are as follows: The plaintiff, Bai, and defendant Wang registered their marriage on January 16, 2017,

**Figure A4.** *Cont.*

and divorced on March 20, 2023. Beginning in September 2022, Wang and defendant Zhang began an extramarital relationship, and between September 2022 and March 20, 2023, Wang transferred a total of 561,766.12 RMB to Zhang through WeChat, Alipay, bank transfers, cash, and shopping. As of the date the plaintiff filed the lawsuit, Zhang had returned 226,090 RMB, leaving 338,676.12 RMB unpaid. The plaintiff argues that during their marriage, Wang made large donations from the marital property to Zhang without the plaintiff's consent, which harmed the plaintiff's legal rights, thus leading to this lawsuit.

Defendant Zhang argues: 1. Defendant Wang concealed her marriage and deceived Zhang, causing serious damage to his reputation and mental well-being, and claims 50,000 RMB for mental damages; 2. The amounts Wang transferred to Zhang are incorrect, including 9,980 RMB, 3,500 RMB, and 3,000 RMB, which Zhang did not receive and should be deducted; 3. Zhang also transferred 127,940 RMB to Wang through Wang Zhen, in addition to the 226,090 RMB mentioned in the plaintiff's complaint; 4. The 5,300 RMB transfer to Zhang on January 11, 2023, was for fireworks purchased during the New Year, which Zhang and Wang jointly set off, so Zhang should not be required to return this money. Additionally, a land lease of 5,000 RMB in Zhang's village in October 2022 should also be deducted. Zhang agrees to return the jewelry, clothes, robot, and steamer gifted by Wang but values them at 19,171 RMB. The donations labeled

**Figure A4.** *Cont.*

with special terms such as "5200" and "1314" totaling 24,315.12 RMB should not be returned. Expenses Zhang incurred for buying shoes (723 RMB), jade (10,000 RMB), and rent (1,000 RMB) should be deducted.

Defendant Wang argues: Zhang knew about her marriage but kept asking Wang for money. The 127,940 RMB came from Wang's client and was transferred through Zhang's account to Wang Zhen’s account, and Wang later sent 30,000 RMB and 60,000 RMB to Zhang. After the completion of Wang's Lingqiu project, equipment was stored at Zhang's home, and no land was leased. Wang acknowledges spending 5,300 RMB on fireworks, which were jointly set off. Regarding the items Wang bought for Zhang, including jewelry, clothes, a robot, and a steamer, she admits to buying shoes and jade for Zhang, but not at the amount claimed. Wang also acknowledges the mistake regarding Zhang's mental distress and sent 20,000 RMB to Zhang after their last meeting in Lingqiu.

The plaintiff presented evidence in support of their claims, including WeChat chat logs, transaction receipts, and shopping records, demonstrating that between September 2022 and March 20, 2023, Wang transferred a total of 564,946.12 RMB to Zhang. Zhang returned 226,090 RMB, leaving 338,676.12 RMB still unpaid. Zhang disputed some amounts, such as 9,980 RMB, 3,000 RMB, and 3,500 RMB, claiming they were not received. These amounts were related to gifts of fruit and dried goods, which the court could not confirm were given to

**Figure A4.** *Cont.*

Zhang, so these amounts were deducted. The cost of shoes (723 RMB) and jade (10,000 RMB), which Zhang bought for Wang, was also deducted. Gifts, including jewelry, clothes, and a robot, were accepted by Zhang but should be returned in their original form. Zhang's request to deduct the rent and other lease-related expenses was not supported by evidence. The fireworks expenses were also to be deducted since both Wang and Zhang participated in the activity.

The court concluded that during the marriage, both spouses had equal rights to manage jointly owned property. Significant decisions on property that are not for daily living needs require mutual agreement. Wang's donation of marital property to Zhang was not for daily needs and was made without the plaintiff's consent, violating the principle of fairness in civil law. Therefore, this donation was invalid. As the marriage ended on March 20, 2023, and there was no explicit agreement on the division of property, the court ruled that Bai is entitled to half of the jointly owned property, amounting to 143,591.06 RMB.

Therefore, the court rules as follows: 1. Defendant Zhang must return 143,591.06 RMB to the plaintiff Bai. 2. Defendant Zhang must return the gifted items, including a gold bracelet, wool fur coat, down jacket, parka, cashmere coat, a robot vacuum, and a steamer. (To be returned within ten days after the judgment becomes effective.) If the payment is not made on time, the defendant must pay double the interest for the

**Figure A4.** *Cont.*

delay, according to the Civil Procedure Law of the People's Republic of China.

The case acceptance fee of 2,213 RMB is to be borne by plaintiff Bai (737.6 RMB), defendant Zhang (737.6 RMB), and defendant Wang (737.8 RMB).

If dissatisfied with this judgment, an appeal can be filed with the Intermediate People's Court of Datong City, Shanxi Province, within fifteen days of receiving the judgment.

Judge: Wang Sheng

June 26, 2023

Clerk: Miao Yuanli

**Figure A4.** A judicial document.

## References


1. Quaresma, P.; Rodrigues, I. A question-answering system for portuguese juridical documents. In Proceedings of the 10th International Conference on Artificial Intelligence and Law, Bologna, Italy, 6–11 June 2005; ser. ICAIL '05; Association for Computing Machinery: New York, NY, USA, 2005; pp. 256–257. [CrossRef]
2. Maxwell, K.T.; Schafer, B. Concept and context in legal information retrieval. In Proceedings of the Legal Knowledge and Information Systems—JURIX 2008: The Twenty-First Annual Conference on Legal Knowledge and Information Systems, Florence, Italy, 10–13 December 2008.
3. Bennett, Z.; Russell-Rose, T.; Farmer, K. A scalable approach to legal question answering. In Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law, London, UK, 12–16 June 2017; ser. ICAIL '17; Association for Computing Machinery: New York, NY, USA, 2017; pp. 269–270. [CrossRef]
4. Šavelka, J.; Ashley, K.D. Legal information retrieval for understanding statutory terms. In *Artificial Intelligence and Law*; Springer: Berlin/Heidelberg, Germany, 2022; Volume 1, pp. 1–45.
5. Lima, J.A.O. Unlocking legal knowledge with multi-layered embedding-based retrieval. *arXiv* **2024**, arXiv:2411.07739.
6. Huang, W.; Jiang, J.; Qu, Q.; Yang, M. Aila: A question answering system in the legal domain. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 5258–5260.
7. 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)*; Burstein, J., Doran, C., Solorio, T.M. Eds.; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. Available online: https://aclanthology.org/N19-1423 (accessed on 3 May 2024).
8. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
9. Paul, S.; Mandal, A.; Goyal, P.; Ghosh, S. Pre-trained language models for the legal domain: A case study on indian law. *arXiv* **2023**, arXiv:2209.06049.
10. Korinek, A. Generative AI for economic research: Use cases and implications for economists. *J. Econ. Lit.* **2023**, *61*, 1281–1317. Available online: https://www.aeaweb.org/articles?id=10.1257/jel.20231736 (accessed on 5 May 2024). [CrossRef]
11. Hang, C.N.; Tan, C.W.; Yu, P.D. MCQGen: A large language model-driven MCQ generator for personalized learning. *IEEE Access* **2024**, *12*, 102261–102273. [CrossRef]
12. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *J. Med. Syst.* **2023**, *47*, 33. [CrossRef]

13. Cui, J.; Ning, M.; Li, Z.; Chen, B.; Yan, Y.; Li, H.; Ling, B.; Tian, Y.; Yuan, L. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv* **2024**, arXiv:2306.16092.
14. Zhou, Z.; Shi, J.-X.; Song, P.-X.; Yang, X.-W.; Jin, Y.-X.; Guo, L.-Z.; Li, Y.-F. LawGPT: A Chinese legal knowledge-enhanced large language model. *arXiv* **2024**, arXiv:2406.04614.
15. Li, H. Lexilaw: A Legal Text Generation System. 2024. Available online: https://github.com/CSHaitao/LexiLaw?tab=readme-ov-file (accessed on 5 November 2024).
16. Huang, Q.; Tao, M.; Zhang, C.; An, Z.; Jiang, C.; Chen, Z.; Wu, Z.; Feng, Y. Lawyer Llama Technical Report. *arXiv* **2023**, arXiv:2406.04614.
17. Branting, L.K.; Pfeifer, C.; Brown, B.; Ferro, L.; Aberdeen, J.; Weiss, B.; Pfaff, M.; Liao, B. Scalable and explainable legal prediction, *Artif. Intell. Law* **2021**, *29*, 213–238. [CrossRef]
18. Mikolov, T.; Yih, W.-T.; Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 746–751.
19. Vale, D.; El-Sharif, A.; Ali, M. Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI Ethics* **2022**, *2*, 815–826. [CrossRef]
20. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
21. Zhang, Y.; Song, K.; Sun, Y.; Tan, S.; Udell, M. Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *arXiv* **2019**, arXiv:1904.12991.
22. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
23. Schwab, P.; Karlen, W. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. *arXiv* **2019**, arXiv:1910.12336.
24. Lundberg, S. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874.
25. Kim, S.; Joo, S.J.; Kim, D.; Jang, J.; Ye, S.; Shin, J.; Seo, M. The CoT Collection: Improving Zero-shot and Few-shot Learning of Language Models via Chain-of-Thought Fine-Tuning. *arXiv* **2023**, arXiv:2305.14045.
26. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 24 824–24 837. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf (accessed on 20 May 2024).
27. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; K′uttler, H.; Lewis, M.; Yih, W.-T.; Rockt′aschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 9459–9474. Available online: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf (accessed on 15 May 2024).
28. Mansurova, A.; Mansurova, A.; Nugumanova, A. QA-RAG: Exploring LLM reliance on external knowledge. *Big Data Cogn. Comput.* **2024**, *8*, 115. Available online: https://www.mdpi.com/2504-2289/8/9/115 (accessed on 14 January 2025). [CrossRef]
29. Lialin, V.; Deshpande, V.; Rumshisky, A. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv* **2023**, arXiv:2303.15647.
30. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
31. Li, X.L.; Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv* **2021**, arXiv:2101.00190.
32. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for Nlp. *arXiv* **2019**, arXiv:1902.00751.
33. Lester, B.; Al-Rfou, R.; Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv* **2021**, arXiv:2104.08691.
34. Liu, X.; Ji, K.; Fu, Y.; Tam, W.L.; Du, Z.; Yang, Z.; Tang, J. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv* **2022**, arXiv:2110.07602.
35. Bai, Y.; Du, X.; Liang, Y.; Jin, Y.; Liu, Z.; Zhou, J.; Zheng, T.; Zhang, X.; Ma, N.; Wang, Z.; et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv* **2024**, arXiv:2403.18058.
36. Yue, S.; Chen, W.; Wang, S.; Li, B.; Shen, C.; Liu, S.; Zhou, Y.; Xiao, Y.; Yun, S.; Huang, X.; et al. DISC-LawLLM: Fine-tuning large language models for intelligent legal services. *arXiv* **2023**, arXiv:2309.11325.
37. Fei, Z.; Shen, X.; Zhu, D.; Zhou, F.; Han, Z.; Zhang, S.; Chen, K.; Shen, Z.; Ge, J. Lawbench: Benchmarking legal knowledge of large language models. *arXiv* **2023**, arXiv:2309.16289.

38. Soffer, S.; Glicksberg, B.S.; Kovatch, P.; Efros, O.; Freeman, R.; Charney, A.W.; Nadkarni, G.N.; Klang, E. A scalable framework for benchmarking embedding models for semantic medical tasks. *medRxiv* **2024**. [CrossRef]
39. Rahutomo, F.; Kitasuka, T.; Aritsugi, M. Semantic cosine similarity. In Proceedings of the 7th International Student Conference on Advanced Science and Technology (ICAST), Mataram, Indonesia, 14 November 2022; University of Seoul: Seoul, Republic of Korea, 2012; Volume 4, p. 1.
40. Lai, J.; Gan, W.; Wu, J.; Qi, Z.; Yu, P.S. Large language models in law: A survey. *arXiv* **2023**, arXiv:2312.03718. [CrossRef]