Elevating Legal LLM Responses: Harnessing Trainable Logical Structures and Semantic Knowledge with Legal Reasoning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved impressive results across numerous domains, yet they experience notable deficiencies in legal question-answering tasks. LLMs often generate generalized responses that lack the logical specificity required for expert legal advice and are prone to hallucination, providing answers that appear correct but are unreliable. Retrieval-Augmented Generation (RAG) techniques offer partial solutions to address this challenge, but existing approaches typically focus only on semantic similarity, neglecting the logical structure essential to legal reasoning. In this paper, we propose the Logical-Semantic Integration Model (LSIM), a novel supervised framework that bridges semantic and logical coherence. LSIM comprises three components: reinforcement learning predicts a structured fact-rule chain for each question, a trainable Deep Structured Semantic Model (DSSM) retrieves the most relevant candidate questions by integrating semantic and logical features, and in-context learning generates the final answer using the retrieved content. Our experiments on a real-world legal QA dataset-validated through both automated metrics and human evaluation-demonstrate that LSIM significantly enhances accuracy and reliability compared to existing methods[1].

1 Introduction

People are inseparable from the protection and constraints of laws, and inevitably faces legal issues in the daily life. However, when confronting legal issues, due to a lack of professional legal knowledge and unaffordable lawyer fees, many people are forced to forgo defending their legal rights (Mansfield and Trubek, 2011; Brescia et al., 2014; Knake, 2012). Although the internet provides a vast array of legal information resources, the uneven quality in terms of authenticity, reliability, and comprehensibility makes it difficult for non-professions to obtain useful legal knowledge (Duranti and Rogers, 2012). Consequently, an expert, understandable, and low-cost legal question-and-answer (Q&A) service is urgently needed. The rapid advancement of Large Language Models (LLMs) has provided ordinary individuals with a way to access affordable legal services, substantially expanding their opportunities to obtain legal assistance (Cheong et al., 2024; Louis et al., 2024). However, given the diversity, complexity, and vague expressions of legal questions, responses generated by LLMs often lack logical specificity and may include hallucinations, exposing users to potential legal risks and financial losses (Dahl et al., 2024). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Li et al., 2024) has emerged as a promising approach to mitigate hallucinations and improve response accuracy by retrieving relevant legal cases and provisions as contextual knowledge sources for LLMs. Nevertheless, existing RAG methods primarily focus on semantic similarity and frequently overlook the intricate logical structures and reasoning essential for addressing complex legal issues, limiting their effectiveness in real-world legal applications.

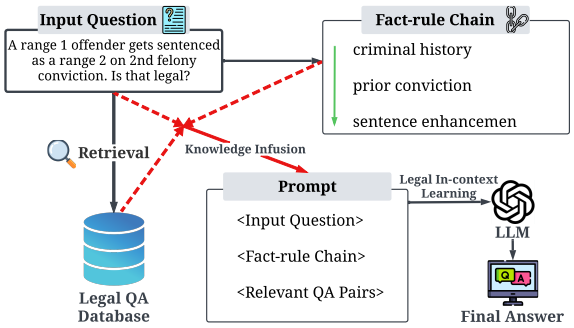

Figure 1: An illustration of our model.

As illustrated in Fig. 1, this paper introduces a novel LLM framework specifically designed to address complex legal QA tasks. Our approach

[1]All data and code will be publicly available, a copy is attached with this submission.

integrates learnable chain-of-thought (CoT) reasoning as logical structures with supervised RAG, enhanced by the Deep Structured Semantic Model (DSSM) (Huang et al., 2013). By embedding logical reasoning into the retrieval process, the framework ensures that both semantic relevance and logical coherence are maintained. Additionally, in-context learning is leveraged to synthesize high-quality answers, utilizing the retrieved auxiliary information to deliver precise, contextually appropriate responses.

The contributions of this work are fourfold:

- We propose a novel LSIM framework, which consists of three components: learnable fact-rule chain, supervised DSSM-powered RAG, and legal in-context learning for precise legal answer generation.
- We employ reinforcement learning to estimate the logical structure of users' legal questions, which integrates both logical structures and semantic information, and navigates LLMs to generate responses like legal professional.
- We extract fact-rule information in the form of chain of thought from users' legal questions, identifying essential facts and applicable legal rules. This enables the system to understand complex legal issues with precision, facilitating the retrieval of highly relevant case law.
- We conducted extensive experiments on a real-world legal QA dataset collected specifically for this study. The results validate the effectiveness and reliability of the proposed framework. To foster future research and development in the legal QA domain, we will release the dataset, code, and models to the community.

2 Related work

2.1 Retrieval-Augmented Generation (RAG)

RAG can significantly improve the model performance by leveraging additional knowledge and has been widely applied in various tasks, such as question & answering (Q&A) (Lewis et al., 2020; Mao et al., 2020), machine translation (Gu et al., 2018), and summarization (Liu et al., 2020; Parvez et al., 2021). With the emergence of LLMs such as LLaMA and ChatGPT, the integration of RAG with LLMs has gained significant popularity and led to significant advancements in multiple tasks (Liu et al., 2023a; Kim et al., 2023; Sharma et al., 2024; Feng et al., 2024).

RAG is also widely applied in research within the legal domain, such as legal Q&A (Cui et al., 2023; Louis et al., 2024; Wiratunga et al., 2024), legal judgment prediction (Wu et al., 2023), legal text evaluation (Ryu et al., 2023), and terminology drafting for legislative documents (Chouhan and Gertz, 2024).

However, most prior research primarily concentrates on improving the performance of retrieval models from a semantic perspective. While semantic information is undoubtedly important, the significance of logical structure is particularly prominent in dealing with legal questions. Legal reasoning often relies on a well-defined logical flow. To address this challenge, our study emphasizes the integration of both semantic information and logical structure in retrieval processing.

2.2 Question & Answering (Q&A)

Q&A is an active research area in NLP that aims to develop systems capable of providing accurate and relevant answers to questions posed in natural language by users based on large knowledge sources (Rogers et al., 2023). Current Q&A studies mainly focus on 1) knowledge retrieval which aims to develop effective and efficient methods to retrieve relevant information from large knowledge bases or corpora (Karpukhin et al., 2020), 2) reading comprehension which aims to build models that can comprehend passages to identify answer-relevant information (Baradaran et al., 2022), 3) multi-hop reasoning, which aims to perform multi-step reasoning by combining information from multiple sources (Wang et al., 2022), and 4) explainable Q&A which aims to generate human-understandable explanations or rationales to support their answers (Latcinnik and Berant, 2020).

2.3 AI Applications in Law

The legal domain has seen increasing interest in applying AI and machine learning techniques to assist with various tasks in Law. One active area of research is using NLP for legal document analysis and information extraction (Zhong et al., 2020). Mistica et al. (2020) created a schema based on related information that legal professionals seek within judgements and performed classification based on it. Sun et al. (2023) proposed a model-agnostic causal learning framework to for legal case matching. There is also work on using AI for legal

judgment prediction, as in Liu et al. (2023b) who develop a neural framework to predict judgments from fact descriptions.

Another emerging application is using AI for legal reasoning and argument mining from texts. Mumford et al. (2023) establihsed a new dataset and explored neural methods to capture patterns of reasoning over legal texts. Zhang et al. (2023) investigated extracting argumentative components like claims and evidence from legal cases. Some researchers are also exploring constitutionality analysis, with Sert et al. (2022) proposing an AI system to predict decisions of the Turkish constitutional court. While promising, these AI-based legal methods still face challenges around interpretability, generalization, and capturing the nuanced reasoning required in law.

3 Methodology

In this section, we propose our Logical-Semantic Integration Model (LSIM), as depicted in Fig. 2.

3.1 Learnable fact-rule chain

In the legal domain, judging a case requires comprehensive consideration of the facts of the case and relevant legal rules. Each judgment process is a reasoning process that needs to combine the facts of the case with legal rules to reach a final conclusion. The user's question and the lawyer's response can be viewed as a complete legal case. Constructing a fact-rule chain for the case helps clarify the entire logical structure and more clearly identify the core issues for the case. Therefore, in this study, we represent the logical structure as the fact-rule chain of the question and its answer. Each node in the fact-rule chain consists of either a fact or a rule. Fact nodes correspond to the specific circumstances of the case, such as illegal search, arrest, and indictment. Rule nodes correspond to the relevant legal basis applicable to legal case circumstances, such as Fourth Amendment, harmless error doctrine, and Federal Rules of Evidence. The complete fact-rule chain illustrates the comprehensive reasoning process for the legal question/answer.

3.1.1 Logical structure extraction

Following Wu et al. (2024), a fact-rule graph $\mathcal{G}$ is constructed using the LLM. Assume our training set $T = \{(q_i, a_i)\}_{i=1}^{N}$ contains N instances, where q_i is the i-th question, and a_i is the real lawyer's answer to q_i. For each question-answer pair (q_i, a_i), the LLM, guided by a tailored prompt, is employed to extract the most relevant fact-rule chain from graph $\mathcal{G}$. The prompt is provided in Appendix A.1[2]. Then, the fact-rule chain C_{q_i} for question q_i is obtained, and $C_{q_i} = \{c_{q_i,1}, c_{q_i,2}, ..., c_{q_i,t}\}$ where $c_{q_i,t}$ is the t-th fact-rule chain node of the question q_i. Similarly, the fact-rule chain C_{a_i} for answer a_i is obtained, and $C_{a_i} = \{c_{a_i,1}, c_{a_i,2}, ..., c_{a_i,t}\}$ where $c_{a_i,t}$ is the t-th fact-rule chain node of the answer a_i[3]. Consequently, the fact-rule chain for all questions C_Q and the fact-rule chain for all answers C_A in the training set are obtained, where $C_Q = \{C_{q_i}\}_{i=1}^{N}$ and $C_A = \{C_{a_i}\}_{i=1}^{N}$[4]. Finally, an average of 2.86 nodes can be extracted from a legal question/answer.

3.1.2 Learning to predict fact-rule chain

The prediction of fact-rule chain is inherently a cumulative and continuous process. Therefore, we consider the fact-rule chain prediction task as a sequential decision-making process, and a reinforcement learning-based approach is empolyed. Given the fact-rule chain C_{q_i} for the legal question q_i, we first encode C_{q_i} using BERT (Kenton and Toutanova, 2019) to obtain its embedding representation:

$$h_{C_{q_i}} = Encode(C_{q_i}). \tag{1}$$

Then, we utilize a policy network $\pi_\theta(n_t|s_t)$ to predict the fact-rule chain for answer a_i step by step, where s_t represents the current state at time step t, and n_t denotes the action (next fact-rule chain node) predicted by the policy network. The initial fact-rule chain $C_{q_i}^{t=0}$ is set to Cq_i, and the initial state s_0 is set to $h_{C_{q_i}}$. At step t, the policy network selects an action n_t based on the current state s_t. Then, the selected node n_t is appended to the current fact-rule chain:

$$C_{q_i}^{t+1} = [C_{q_i}^{t}, n_t], \quad n_t \sim \pi_\theta(n_t|s_t) \tag{2}$$

Subsequently, the state embedding is updated using the new fact-rule chain:

$$s_{t+1} = Encode(C_{q_i}^{t+1}). \tag{3}$$

[2]In cases where an exact fact/rule node cannot be found, the most similar node is selected.

[3]Based on our analysis and statistics, we have identified that a maximum of four key elements are sufficient to cover the main content of a legal question/answer. Therefore, in this paper, the value of t ranges from 1 to 4.

[4]C_Q is a set composed of fact-rule chains for N legal questions, and C_A is a set composed of fact-rule chains for N legal answers.

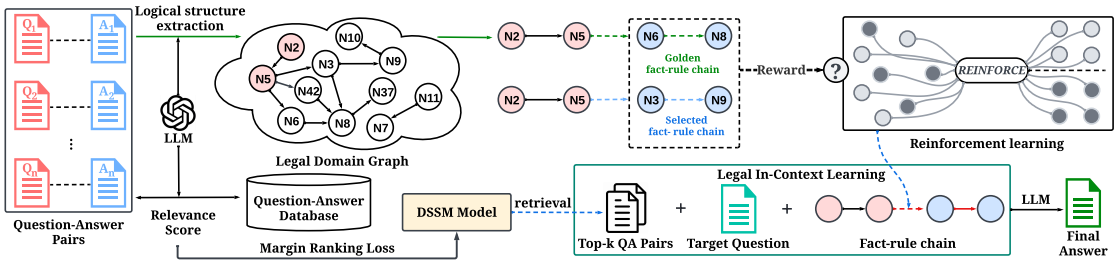

Figure 2: The overall framework of LSIM. The network consists of three modules: learnable fact-rule chain, supervised DSSM-powered RAG, and legal in-context learning.

This process is repeated until a maximum number of steps is reached or no valid next node can be selected. The policy network is implemented as a multi-layer perceptron (MLP). The REINFORCE algorithm (Williams, 1992) is employed to train the policy network, which is a classic policy gradient method in reinforcement learning. The REINFORCE algorithm focuses strongly on maximizing long-term rewards. This enables the model to consider more far-reaching goals and impacts, rather than merely immediate predictive performance. In complex tasks such as legal reasoning, each thought step is built upon the foundation of previous logical reasoning. The REINFORCE algorithm can effectively simulate this process, learning the logical relationships between nodes in the fact-rule chain, thereby ensuring that the generated fact-rule chain is logically coherent and accumulative. The training objective is to maximize the expected cumulative reward:

$$J(\theta) = \mathbb{E}\pi_\theta[\sum_{t=0}^{T} r_t], \tag{4}$$

where r_t is the reward at step t, and T is the maximum number of steps. The reward r_t is defined as follows:

$$r_t = \begin{cases} 1, & \text{if } n_t \in C_{a_i} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where C_{a_i} is the ground-truth fact-rule chain for answer a_i.

During inference, the trained policy network is employed to predict the fact-rule chain for a given legal question. Assume the inference step is z, the predicted fact-rule chain is $C^z_{q_i}$ for question q_i, and $C^z_{q_i}$ is the predicted logical structure. Appendix B provides examples illustrating the fact-rule chain results.

3.2 Supervised DSSM-powered RAG

Deep Structured Semantic Model (DSSM) (Huang et al., 2013) is a used to retrieve the most relevant questions to the current user's question q_i from the database. Let D be the database of candidate questions, and $D = \{(q^D_j, a^D_j)\}^M_{j=1}$ contains M instances, where q^D_j is the j-th candidate question in D, and a^D_j is the real lawyer's answer to q^D_j.

Given a legal question $q_i \in T$, its logical structure $C^z_{q_i}$ can be obtained by the method described in Section 3.1.2. Similarly, for each candidate question $q^D_j \in D$, its logical structure $C^z_{q^D_j}$ can also be obtained. Then we encode each of them independently using the same encoder:

$$\begin{aligned} h_{q_i} &= Encode(q_i), \\ h_{C_{q_i}} &= Encode(C^z_{q_i}), \\ h_{q^D_j} &= Encode(q^D_j), \\ h_{C_{q^D_j}} &= Encode(C^z_{q^D_j}). \end{aligned} \tag{6}$$

Subsequently, $h_{C_{q_i}}$, which represents the logical structure features, and h_{q_i}, which represents the semantic features, are concatenated together to form the features for the current question q_i:

$$e_{q_i} = [h_{C_{q_i}}, h_{q_i}]. \tag{7}$$

Similarly, the features for candidate question q^D_j can be obtained:

$$e_{q^D_j} = [h_{C_{q^D_j}}, h_{q^D_j}]. \tag{8}$$

The DSSM model is composed of a multi-layer perceptron (MLP) and computes a relevance score p_{ij} between q_i and candidate question q^D_j:

$$\begin{aligned} x_1 &= \tanh(W_1[e_{q_i}, e_{q^D_j}] + b_1) \\ x_2 &= \tanh(W_2 x_1 + b_2) \\ x_3 &= \tanh(W_3 x_2 + b_3) \\ p_{ij} &= W_4 x_3 + b_4, \end{aligned} \tag{9}$$

where W_1, W_2, W_3, and W_4 are weights, and b_1, b_2, b_3 , and b_4 are bias. The margin ranking loss is employed, which encourages the model to assign higher scores to more relevant cases. For each question q_i, we select the candidate question in the database with the highest annotated relevance score as the positive example c_i^+, and the candidate question with the lowest score as the negative example c_i^-. The annotated relevance scores are generated by the LLM. Specifically, the LLM evaluates the relevance between the current query and each candidate question. These relevance scores are assigned on a scale from 0 to 5, where a score of 0 indicates minimal relevance and a score of 5 denotes the highest level of similarity.The prompt is provided in Appendix A.2 The loss function is defined as:

$$\mathcal{L}(q_i, c_i^+, c_i^-) = \max(0, \alpha - p(q_i, c_i^+) + p(q_i, c_i^-)), \tag{10}$$

where α is a hyperparameter.

During inference, for each question q_i, we compute the relevance scores between q_i and all candidate questions in the database D using the trained DSSM model. The top-K candidate questions with the highest scores are the final retrieval results.

3.3 Legal in-context learning

After retrieving the top-K most relevant questions $q_{j_1}^D, q_{j_2}^D, ..., q_{j_K}^D$ from the database D for the current question q_i, we concatenate them with their corresponding answers $a_{j_1}^D, a_{j_2}^D, ..., a_{j_K}^D$ to form the context for in-context learning:

$$context_i = [(q_{j_1}^D, a_{j_1}^D), (q_{j_2}^D, a_{j_2}^D), ..., (q_{j_K}^D, a_{j_K}^D)]. \tag{11}$$

This context provides the LLM with relevant examples of how relevant legal questions have been answered by real lawyers in the past. Following Wu et al. (2023) and considering the length limit of the prompt, K is set to 3 in our experiments.

Then, the current question q_i, the logical structure C_{q_i}, and $context_i$ are provided to the LLM to generate an answer. The prompt is provided in Appendix A.3:

$$a_i' = LLM(q_i, C_{q_i}, context_i). \tag{12}$$

4 Experiments

4.1 Datasets

We use real-world legal question and answer (Q&A) data collected from JUSTIA[5]. The dataset comprises 16,190 legal questions posed by users in the field of criminal law, with each question receiving responses from an average of 1.26 lawyers. The dataset covers legal questions from all 50 states in the USA. The average length of the questions is 67 words, and the responses average 40 words. In the experiment, the total dataset comprising 16,190 samples is divided using an 8:2 ratio: 80% of the data (12,952 samples) is used as the database for retrieval, and 20% (3,283 samples) is used for training and testing. Subsequently, the 3,238 samples are further divided into training and testing sets using the same 8:2 ratio. The specific information is presented in Table 1. Appendix C presents some samples of the collected data.

Type	Number
Database	12,952
Training	2590
Testing	648

Table 1: Statistics of data we collected.

4.2 Baselines and evaluation metrics

Baselines. We implement the following baselines for comparison: **BM25** (Robertson and Walker, 1994), a classic bag-of-words information retrieval model, is used to retrieve the question from the database that most closely matches the user's query. **Bert** (Kenton and Toutanova, 2019), **Roberta** (Liu et al., 2019), **Sentence T5** (Ni et al., 2022a), **INSTRUCTOR** (Su et al., 2023), **GTR** (Ni et al., 2022b), **BGE-m3** (Chen et al., 2024), **text-embedding-ada-002**, **text-embedding-3-small**, and **text-embedding-3-large**[6], are employed to generate embeddings for the user's query and questions in the database. Similarity calculations are then used to determine the closest match. Following Louis et al. (2024); Wu et al. (2023, 2024), **LLaMA-2-13B**, **LLaMA-3-8B**[7], and **GPT-4o**, serve as the LLM baselines in our study. They can generate responses to the posed questions.

Evaluation metrics. To evaluate our model, both automatic and human evaluations are used. For automatic evaluation, the commonly used text generation metrics, **ROUGE** (ROUGE-1, ROUGE-2, and ROUGE-L) (Lin, 2004), **METEOR** (Banerjee and Lavie, 2005), and **BERTScore** (Zhang

[5]https://www.justia.com/

[6]They are advanced embedding models developed by OpenAI.

[7]https://llama.meta.com/

Method	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
LLaMA-2-13B w/o RAG	17.09	13.13	1.91	12.09	81.24
BM25	17.75	13.98	2.22	12.81	81.10
Bert-Base	17.96	13.94	2.18	12.69	81.33
Roberta	17.87	13.85	2.21	12.71	81.27
SentenceT5-Base	17.86	13.99	2.28	12.85	81.74
INSTRUCTOR-Base	17.92	14.12	2.26	12.82	81.81
GTR-Base	17.76	13.95	2.22	12.81	81.58
BGE-m3	17.85	13.75	2.14	12.59	81.35
text-embedding-ada-002	17.76	13.89	2.22	12.70	81.45
text-embedding-3-small	17.89	14.02	2.16	12.76	81.53
text-embedding-3-large	18.03	14.13	2.23	12.93	81.62
LSIM	**20.55**	**16.10**	**2.58**	**14.52**	**83.12**

Table 2: Performance on legal response generation using LLaMA-2-13B (%).

Method	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
LLaMA-3-8B w/o RAG	17.13	11.56	1.69	10.62	81.91
BM25	17.84	13.34	2.09	12.10	82.47
Bert-Base	17.68	13.13	1.99	11.92	82.53
Roberta	18.03	13.38	2.14	12.15	82.56
SentenceT5-Base	18.40	13.58	2.16	12.40	82.50
INSTRUCTOR-Base	18.39	13.61	2.28	12.29	82.49
GTR-Base	18.10	13.54	2.15	12.32	82.48
BGE-m3	18.08	13.46	2.08	12.12	82.52
text-embedding-ada-002	17.87	13.24	2.04	12.01	82.48
text-embedding-3-small	18.32	13.70	2.22	12.44	82.49
text-embedding-3-large	18.62	13.82	2.24	12.53	82.52
LSIM	**21.00**	**16.30**	**2.63**	**14.74**	**83.23**

Table 3: Performance on legal response generation using LLaMA-3-8B (%).

et al., 2019) are employed. Human evaluation focuses on three aspects: 1) **Accuracy**: the aspect evaluates whether the generated answers are correct and free from factual errors. 2) **Specificity**: this aspect measures whether the responses are directly related to the specific issues raised in the question, providing clear and targeted answers rather than generalized responses. 3) **Adoptability**: this aspect assesses whether the responses generated by the model are practically useful and can be directly provided to users. Three legal professionals were invited to evaluate the answers generated by LLaMA-3-8B, text-embedding-3-large, and our proposed Model LSIM. Each dimension is rated on a scale of 1-5, with 5 being the highest score.

4.3 Experiment Settings

For GPT-4o, the sampling parameters are set with a temperature of 0.8. For LLaMA-2-13B and LLaMA-3-8B, the sampling parameters are set with a temperature of 0.8 and a top-p value of 0.9. The maximum token limit per generation is set at 4096. For the LSIM method, the word embeddings are initialized using BERT. Adam is used as the optimizer. The learning rates for the policy network and the DSSM are both set to 1e-4. The number of epochs for the policy network and the DSSM are set to 30 and 50, respectively. All the results are the average values of three repeated runs.

4.4 Experiment Results

Tables 2, 3, and 4 present the experimental results obtained when LLaMA-2-13B , LLaMA-3-8B, and GPT-4o are used as base architectures, respectively. It can be observed that LLMs struggle to effectively answer legal questions when lacking support from relevant cases. When provided with relevant cases as auxiliary information, the per-

Method	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
GPT-4o w/o RAG	17.92	12.36	2.20	11.46	81.60
BM25	18.41	12.88	2.22	11.84	82.01
Bert-Base	18.53	12.90	2.24	11.91	82.11
Roberta	18.50	13.01	2.24	11.94	82.16
SentenceT5-Base	18.57	12.97	2.23	11.90	82.05
INSTRUCTOR-Base	18.55	12.92	2.30	11.92	82.04
GTR-Base	18.40	12.84	2.25	11.80	82.00
BGE-m3	18.51	12.96	2.25	11.88	82.05
text-embedding-ada-002	18.35	12.65	2.23	11.73	81.98
text-embedding-3-small	18.47	12.88	2.21	11.86	81.97
text-embedding-3-large	18.69	13.02	2.31	12.01	82.04
LSIM	**21.54**	**14.09**	**2.31**	**12.94**	**82.68**

Table 4: Performance on legal response generation using GPT-4o (%).

User's Question		
Someone I know has been accused of touching a child. She is 18 now and is making these allegations. How can he clear himself. A friend of mine has been accused of inappropriately touching a young child when she was six, that's what she is saying. She is 17 or 18 now. My friend is very upset and is wrongfully accused. Can he take her to court to get his name cleared? If so, what steps should he go about this?		
Response		
Lawyer	**LLaMA-2-13B w/o RAG**	**LSIM**
Depending on the allegations and to whom she is making these claims, this could result in serious criminal charges against your friend. If it were me, I'd keep silent about the allegations. And if he is contacted by law enforcement, definitely get a lawyer before speaking with them.	I understand that your friend is upset and wants to clear their name. However, the appropriate course of action depends on various factors, including the specific circumstances and legal jurisdiction. I recommend that your friend consults with a lawyer who specializes in criminal law to discuss their options and determine the best approach. The lawyer can guide them on how to proceed, potentially including seeking legal action for defamation or slander, and navigating the legal system to clear their name. It's essential to prioritize a thorough and fair investigation, ensuring that all parties involved receive a just outcome.	Your friend should seek legal counsel from a competent attorney who specializes in criminal defense and child abuse cases. The attorney can guide your friend on how to proceed to clear their name and address the allegations. It's essential to remain silent and avoid any contact with the accuser, as any statements or interactions may be used as evidence. The attorney can help your friend understand their rights and options, including the possibility of filing a defamation claim. However, the primary focus should be on addressing the criminal allegations and ensuring your friend's name is cleared.

Figure 3: An illustrative example. The sentences highlighted in red indicate remaining silent, while those highlighted in blue indicate seeking professional legal counsel.

Method	Acc.	Spec.	Adopt.
LLaMA-3-8B	4.08	4.33	4.25
text-embedding-3-large	4.35	4.35	4.41
LSIM	4.65	4.47	4.65

Table 5: Results of human evaluation.

formance of LLMs in answering legal questions is improved. When LLaMA-3-8B is employed as the LLM baseline, our proposed LSIM algorithm achieves the best performance across all metrics. Compared with the LLaMA-3-8B w/o RAG, our proposed LSIM model achieves improvements of 3.87% on METEOR, 4.74% on ROUGE-1, 0.94% on ROUGE-2, 4.12% on ROUGE-L, and 1.32% on BERTScore. Compared to the best performing baseline model text-embedding-3-large, LSIM model achieves improvements of 2.38%, 2.48%, 0.39%, 2.21%, and 0.71% on METEOR, ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore. When LLaMA-2-13B is employed as the LLM baseline, our proposed LSIM algorithm also achieves the best performance across all metrics. Compared with the LLaMA-2-13B w/o RAG, our proposed LSIM model achieves improvements of 3.46% on METEOR, 2.97% on ROUGE-1, 0.67% on ROUGE-2, 2.43% on ROUGE-L, and 1.88% on BERTScore. When GPT-4o is employed, compared to the best performing baseline model text-embedding-3-large, LSIM model achieves improvements of 2.85%, 0.93%, and 0.64% on METEOR, ROUGE-L, and BERTScore, respectively. All results demonstrate that our LSIM algorithm achieves the best performance.

The results of the human evaluation are shown in Table 5. Our LSIM model achieves the best performance in terms of accuracy, specificity, and adoptability. These results highlight the effective-

The most relevant questions retrieved by LSIM
A false allegation of inappropriate touching was made and a polygraph is being requested is there a way to dismiss this? My teen step-daughter has a history of bad behavior and being unruly. During a recent counseling session, she accused me of touching her while giving her a hug last June. CPS is involved and I'm now being asked to take a polygraph. Since the incident, she has been sent to stay with her grandmother after sneaking out and breaking a neighbors window.
Lawyer's Answer
I recommend you keep your mouth shut and do not post anything else online. Hire a competent attorney today to counsel you on your possible criminal charges and how to conduct yourself during this DCS and/or LEO investigation. Again do not talk to anyone about this and have no contact with the girl.

Table 6: The most relevant questions retrieved by LSIM.

Method	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
LSIM	**21.00**	**16.30**	**2.63**	**14.74**	**83.23**
LSIM w/o LS	19.10	14.18	2.22	12.88	82.45
LSIM w/o SI	18.77	13.83	2.17	12.62	82.48

Table 7: Ablation study when using LLaMA-3-8B as the LLM architecture (%).

ness of our proposed LSIM model.

4.5 Case Study

Figure 3 presents a comparison of LLaMA-2-13B directly answering a legal question (LLaMA-2-13B w/o RAG) and utilizing our LSIM framework to respond to the legal question. For this given question, there are two main points in a real lawyer's response: keep silent and get a lawyer. However, the response generated by LLaMA-2-13B w/o RAG is comparatively generic, merely mentioning seeking legal counsel. By leveraging our LSIM model, we retrieve the three most relevant questions related to the given query. Tabel 6 presents the most relevant question retrieved by LSIM. The answer to this question advises the user to "keep your mouth shut" and "hire a competent attorney".

By incorporating insights from the retrieved relevant questions, the LSIM model generates a response that covers two crucial aspects: remain silent and seek legal counsel. These key points align closely with the advice given by the real lawyer. The responses generated by our LSIM framework exhibit a higher degree of professionalism and more closely mirror the advice typically provided by a lawyer. More examples are provided in Appendix D.

4.6 Ablation Study

An ablation study is also conducted on the LSIM framework when using LLaMA-3-8B as base LLM. The results are presented in Tabel 7. LSIM w/o LS refers to the LSIM model without the Logical Structure module. LSIM w/o SI refers to the LSIM model without the Semantic Information module. The ablation study demonstrates that both the logical structure (LS) and semantic information (SI) modules contribute positively to the overall performance of the LSIM model. The best results are achieved by the full LSIM model, which combines the effects of both the LS and SI components.

5 Conclusion

This paper addresses the inherent limitations of LLMs in generating professional legal responses. We propose a novel framework, LSIM, designed to enhance the legal LLM reasoning by integrating the learnable logical structure and semantic information of legal questions. The LSIM framework is composed of three key components: (1) Reinforcement learning predicts the fact-rule chain of thought for a given legal query, guiding the reasoning process; (2) a supervised RAG retrieves the most relevant questions by integrating logical and semantic information; and (3) the fact-rule chain, relevant retrieved questions, and their corresponding answers are provided as auxiliary reference information to the LLM, enabling it to generate precise, contextually relevant responses.

We validate the effectiveness of LSIM through experiments on a real-world legal QA dataset. Results from both automated metrics and manual evaluations demonstrate the superior performance of the proposed framework in delivering accurate and expert-level legal answers.

In the future, we plan to extend the LSIM framework to other specialized domains, such as healthcare and finance. Additionally, we aim to incorporate multi-turn interactions with users, leveraging their real-time feedback to further refine the model's performance and adaptability.

Limitations

The effectiveness of the RAG-based model heavily depends on the availability of databases. Consequently, our model's performance may degrade due to the lack of sufficient relevant legal cases to retrieve, which hinders the model's adaptability and utility in regions where legal cases are scarce.

As another limitation, our study is limited to single-turn interactions with LLMs. Expanding to multi-turn interactions could enhance the model's ability to develop a more nuanced and comprehensive understanding of user queries. However, achieving this will require a redesign of the proposed model to effectively manage contextual continuity and iterative feedback across multiple exchanges.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Raymond H Brescia, Walter McCarthy, Ashley McDonald, Kellan Potts, and Cassandra Rivais. 2014. Embracing disruption: How technological change in the delivery of legal services can improve access to justice. *Alb. L. Rev.*, 78:553.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2454–2469.

Ashish Chouhan and Michael Gertz. 2024. Lexdrafter: Terminology drafting for legislative documents using retrieval augmented generation. *arXiv preprint arXiv:2403.16295*.

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

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*.

Luciana Duranti and Corinne Rogers. 2012. Trust in digital records: An increasingly cloudy legal area. *Computer Law & Security Review*, 28(5):522–531.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11661–11665. IEEE.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009.

Renee Newman Knake. 2012. Democratizing the delivery of legal services. *Ohio St. LJ*, 73:1.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. In *arXiv:2004.05569*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. 2023a. Reta-llm: A retrieval-augmented large language model toolkit. *arXiv preprint arXiv:2306.05212*.

Shangqing Liu, Yu Chen, Xiaofei Xie, Jing Kai Siow, and Yang Liu. 2020. Retrieval-augmented generation for code summarization via hybrid gnn. In *International Conference on Learning Representations*.

Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023b. Ml-ljp: Multi-law aware legal judgment prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1023–1034.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22266–22275.

Marsha M Mansfield and Louise G Trubek. 2011. New roles to solve old problems: Lawyering for ordinary people in today's context. *NYL Sch. L. Rev.*, 56:367.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*.

Meladel Mistica, Geordie Z. Zhang, Hui Chia, Kabir Manandhar Shrestha, Rohit Kumar Gupta, Saket Khandelwal, Jeannie Paterson, Timothy Baldwin, and Daniel Beck. 2020. Information extraction from legal documents: A study in the context of common law court judgements. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 98–103.

Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2023. Combining a legal knowledge model with machine learning for reasoning with legal cases. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 167–176.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022a. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. 2022b. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

Md Rizwan Parvez, Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2719–2734.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.

Anna Rogers, PictureMatt Gardner, and PictureIsabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min, and Jy-Yong Sohn. 2023. Retrieval-based evaluation for llms: A case study in korean legal qa. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 132–137.

Mehmet Fatih Sert, Engin Yıldırım, and İrfan Haşlak. 2022. Using artificial intelligence to predict decisions of the turkish constitutional court. *Social Science Computer Review*, 40(6):1416–1435.

Sanat Sharma, David Seunghyun Yoon, Franck Dernoncourt, Dewang Sultania, Karishma Bagga, Mengjiao Zhang, Trung Bui, and Varun Kotte. 2024. Retrieval augmented generation for domain-specific question answering. *arXiv preprint arXiv:2404.14760*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121.

Zhongxiang Sun, Jun Xu, Xiao Zhang, Zhenhua Dong, and Ji-Rong Wen. 2023. Law article-enhanced legal case matching: A causal learning approach. In *Proceedings of ACM SIGIR*, pages 1549–1558.

Yu Wang, Vijay Srinivasan, and Hongxia Jin. 2022. A new concept of knowledge based question answering (KBQA) system for multi-hop reasoning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4007–4017.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering. *arXiv preprint arXiv:2404.04302*.

Yang Wu, Chenghao Wang, Ece Gumusel, and Xiaozhong Liu. 2024. Knowledge-infused legal wisdom: Navigating llm consultation through the lens of diagnostics and positive-unlabeled reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12060–12075.

Gechuan Zhang, Paul Nulty, and David Lillis. 2023. Argument mining with graph representation learning. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, page 371–380.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. In *ACL*, pages 5218–5230.

A Prompt details

A.1 Relevant fact-rule path extraction

"Please select 1 to 4 nodes from the provided graph that are most relevant to the legal question/answer. Ensure that the selected nodes are interconnected."

A.2 Similarity Score

"Please score the similarity of Question2 to Question1, focusing specifically on the events described in each legal question. Rate the similarity of Question2 to Question1 on a scale from 0 to 5, where 0 indicates that Question2 is completely different from Question1, and 5 indicates that Question2 is exactly the same as Question1."

A.3 Generation of the final response

"Your task is to provide legal advice on the user's question. I will provide you with the logical structure of the user's question, along with similar questions previously asked by other users and the responses given by real lawyers. Please use this information to generate a response to the user's question."

B Illustrative examples of the fact-rule chain results

Illustrative examples of the fact-rule chain results are provided in Figure 4.

C Illustrative samples of the collected data

Illustrative samples of the collected data are provided in Table 8.

D Case study

Another case study is provided in Figure 5.

Question	Fact-rule chain of the question	Answer	Fact-rule chain of the answer	Logical structure
A range 1 offender gets sentenced as a range 2 on 2nd felony conviction. Is that legal?	criminal history, prior conviction, sentence enhancement	It is definitely possible. The Sentencing can be appealed but it must be brought up in a Motion For New Trial first. However, the difference is 35% instead of 30% to be served prior to Parole Eligibility, in addition to a larger number of minimum years Sentence.	appeal, sentencing guidelines	criminal history, prior conviction, sentence enhancement, appeal, sentencing guidelines
Can someone use text message as Evidence to against another person in court, Or Investigated Even citizens arrested.	evidence	Yes, but the text messages have to be authenticated and must be relevant. The text messages generally can be authenticated by calling as witness the person who sent it or received the message.	evidence admissibility	evidence, evidence admissibility
I was arrested with a meth pipe only. Is that possession of methamphetamines or possession of Paraphernalia?	evidence, possession	If the government sends the pipe out for testing, and they are able to scrape any meth out of it, they could charge you with possession of methamphetamine. It's possible for this type of a charge to be filed several months after the police took the pipe as they wait for the lab results. It might be a good idea to consult with an attorney so you can prepare yourself for potential charges. If the pipe was clean, then you can't be charged with possession of meth.	controlled substance, criminal law	evidence, possession, controlled substance, criminal law

Figure 4: Illustrative examples of the fact-rule chain results. The red text represents the fact-rule chain for the question, and the green text represents the fact-rule chain for the answer. Together, they form the logical structure.

ID	Location	User's question	Lawyer's answer
1	California	My boyfriend is incarcerated for PC 212.5, in 2017 he was sentenced to 9 years, 667a enhancements added to the sentencing. He has a parole release date in 2023, now after taking many classes and being in fire camp. Is there anyway to reduce more time after new laws passed regarding extra time for enhancements?	Prison prior-667.5(b) enhancements are now gone-but not 667a enhancements. If he successfully participated as a hand crew member at fire camp, once he's been released from custody, he can apply for relief from the convictions that sent him there, and parole, under the new 1203.4b P.C.
2	Florida	My friend legally purchased but did not register a gun and it was found by police in the room she and her Bf share. He is a felon and took the charge because she has a pending DCF case. If she goes back and takes the charge, having no priors, is it possible she would just get probation and not jail time?	What would the charge be, failure to register a firearm? To my knowledge, Florida doesn't require registration, the boyfriend could be charged with being a felon in possession, but I don't think there is any charge she could "take".
3	Florida	I requested an Officer to press charges for my stepson kicking me in the back and vandalizing my home with a sharpie. However, the officer said I did not adopt him and he is allowed to throw things around the house. I am trying to get a hold of a supervisor and I was told to call back when he is back from vacation. Please Advise. I feel strongly the police are bias and not doing there job. Also, my wife is trying to help me with full support of charges. This is odd of the Police.	So, it is definitely not the law that someone is allowed to kick you in the back and vandalize your house just because you didn't adopt him. The crime for vandalizing is called "Criminal Mischief," governed by Florida Statute 806.13. The crime for kicking you in the back is called "Battery," governed by Florida Statute 784.03. Neither crime has a defense or exception having to do with someone's adoption status. Good luck.

Table 8: Illustrative samples of the collected data.

User's Question		
My name is Krista, I'm a victim in a criminal case and I pressed charges on my boyfriend but I want to drop them. He's a convicted felon and was Charged with it kidnapping burglary theft reckless endangerment aggravated assault simple assault and possession of a firearm		
Response		
Lawyer	**LLaMA-2-13B w/o RAG**	**LSIM**
Once charges are filed by the Commonwealth you no longer have the ability to drop them. You become an important witness in the case but the case is the District Attorney's. Those charges sound very serious and very few DAs would simply walk away from that case. Remember their job is to protect everyone, not just you. Be open with your advocate/ADA. They are required to let you have a say but not required to follow your wishes. However, in same cases they will let you help guide the case.	I understand your concerns, Krista. As a victim in a criminal case, you have the right to drop the charges against your boyfriend at any time. However, it's important to understand that the State may still proceed with the case, especially considering the severity of the charges and your boyfriend's prior conviction as a felon. In your situation, since your boyfriend has a history of criminal behavior, it's likely that the prosecutor will want to continue with the case to ensure public safety. Additionally, the State may have already invested significant resources in the investigation and prosecution of the case. If you're still determined to drop the charges, you should contact the prosecutor's office and inform them of your decision. Be prepared for them to ask you why you're dropping the charges, and be honest about your reasons. They may also ask you to provide a written statement explaining your decision. It's also important to note that, even if you drop the charges, your boyfriend may still face consequences for his actions. As a convicted felon, he may be subject to stricter sentencing guidelines if he's found guilty of any future crimes. I would advise you to consult with a lawyer who specializes in criminal law to discuss your options and the potential consequences of dropping the charges. They can help you navigate the process and ensure that your rights are protected throughout.	Unfortunately, as a victim who has already pressed charges against your boyfriend, you may not have the authority to simply "drop" the charges. The prosecution now has control of the case, and it is up to the prosecutor's discretion to decide whether to proceed with the charges. If you want to drop the charges, you will need to speak with the prosecutor's office and explain your reasons for wanting to do so. They may be willing to dismiss the charges, especially considering your boyfriend's criminal history, but this is not guaranteed. It is also important to note that the fact that your boyfriend is a convicted felon may not necessarily influence the prosecutor's decision to drop the charges. The prosecution's primary consideration is typically the strength of the evidence against the defendant, rather than the defendant's criminal history. If you do decide to drop the charges, you should also consider seeking a protective order to protect yourself from your boyfriend. As a victim of domestic violence, you may be eligible for a protective order that can help keep you safe from further harm. Please note that it is always best to consult with a lawyer who specializes in domestic violence cases to discuss your specific situation and the best course of action for you.

Figure 5: Case study.