Towards Expert Legal LLM Responses: Logical Structure and Semantic Information Integration

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated excellent performance across various fields. Nevertheless, they exhibit notable deficiencies when addressing legal questions. In the legal field, LLMs often provide generalized responses, lacking the necessary specificity for expert legal advice. Additionally, they tend to provide answers that appear correct but are unreliable due to issues with hallucination. Retrieval-Augmented Generation (RAG) is a popular approach to addressing these issues. However, existing methods often focus solely on semantic-level similarity, neglecting the logical structure relationships between different legal questions. In this paper, we propose a Logical-Semantic Integration Model (LSIM), which consists of three components. First, reinforcement learning is used to predict the fact-rule chain of thought for the given question. Secondly, the DSSM model that integrates logical structure and semantic information is used to retrieve the most relevant candidate questions from the database. Finally, in-context learning is used to generate the final answer. Experiments on a real-world legal QA dataset, using both automated evaluation metrics and human evaluation, demonstrate the effectiveness of the proposed method. The dataset will be released to the community to promote the development of the legal QA field[1].

1 Introduction

Everyone is inseparable from the protection and constraints of the law, and inevitably faces legal issues that need resolution in the daily lives. However, when confronting legal issues, many individuals lack professional legal knowledge and are forced to forgo defending their legitimate rights and interests. Even those who intend to use legal means to protect their rights often face prohibitive lawyer fees as an obstacle (Mansfield and Trubek, 2011; Brescia et al., 2014; Knake, 2012). Although the modern internet provides a vast array of legal information resources, the quality of this information varies greatly, making it difficult to discern authenticity and obtain reliable legal knowledge (Duranti and Rogers, 2012). Additionally, the specialized language used in legal documents often poses a significant barrier to understanding for non-professionals. Consequently, there is an urgent need for an expert and low-cost legal question-and-answer (Q&A) service to provide individuals with professional legal assistance when facing legal issues, preserving their lawful entitlements. The rapid advancement of Large Language Models (LLMs) has provided ordinary individuals with a way to access affordable legal services, substantially broadening their opportunities to obtain legal assistance (Cheong et al., 2024; Louis et al., 2024). However, given the diversity and complexity of legal questions, responses generated by LLMs are typically verbose and vague, lacking specificity. Moreover, the hallucination problem inherent in LLMs demands careful consideration (Dahl et al., 2024). If LLMs generate responses that appear reasonable on the surface but are actually inconsistent with facts or legal provisions, users may be exposed to legal risks and financial losses. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Li et al., 2024) is an effective method for mitigating LLM hallucinations and enhancing response accuracy. It retrieves cases or legal provisions relevant to the given question and provides them to LLMs as additional knowledge sources. However, current RAG algorithms typically only consider semantic-level similarity. Legal questions often involve complex logical structures, and merely considering semantic relevance may not sufficiently

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

capture the core aspects of the problem.

To further enhance the performance of LLMs in the legal QA task, we propose a novel framework that integrates logical structures and semantic information. Given a user question, first, the LLM is employed to extract a chain-of-thought (CoT), which consists of relevant facts and rules pertaining to the question at hand. Secondly, a reinforcement learning method is utilized to predict the corresponding CoT for the answer, based on the extracted CoT from the user's question. The CoTs for both the question and the answer collectively serve as the logical structure of the user's question. Subsequently, both the logical structure and the semantic information associated with the current question are fed into the Deep Structured Semantic Model (DSSM) (Huang et al., 2013) as input features to retrieve several highly relevant questions from the database. Finally, the retrieved relevant questions and answers, along with the user's question and its logical structure, are provided to LLMs using in-context learning to generate high-quality answers to the current user's question. We collected a real-world legal QA dataset and conducted experiments on it. Both automated evaluation metrics and human evaluations were performed. The experiments demonstrate the effectiveness of our proposed framework.

To sum up, our contributions can be summarized as follows:

- We propose a novel framework named LSIM, which consists of three components: fact-rule chain of though prediction, DSSM retrieval module, and in-context Learning.
- Reinforcement learning utilized to obtain the logical structure of users' questions. Both the logical structure and semantic information are integrated to enhance the ability of LLMs in generating expert legal responses.
- We extract fact-rule information in the form of chain of thought from users' legal questions, capturing essential facts and applicable legal rules. This approach allows for a precise understanding of the core issues and aids in identifying relevant cases.
- Experiments are conducted on a real-world legal QA dataset we collected, and the experimental results validate the effectiveness of our proposed framework. The dataset will be released to the community to promote the development of the legal QA field.

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 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) ex-

plainable 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. (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. (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. (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. (Mumford et al., 2023) establihsed a new dataset and explored neural methods to capture patterns of reasoning over legal texts. Zhang et al. (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. (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 describe our Logical-Semantic Integration Model (LSIM), as shown in Fig. 1.

3.1 Logical Structure Inference

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 CoT 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 CoT of the question and its answer. Each node in the CoT consists of either a fact or a rule. Fact nodes corresponding to the specific circumstances of the case, while rule ones corresponding to the relevant legal basis applicable to those circumstances.

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 is used to extract the most matching fact-rule path in graph $\mathcal{G}$. Then, the chain of thought C_{q_i} for 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 chain of thought node of the question q_i. Similarly, the chain of thought C_{a_i} for 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 chain of thought node of the answer a_i. Consequently, the chain of thought for questions C_Q and the chain of thought for answers C_A in the training set are obtained, where $C_Q = \{C_{q_i}\}_{i=1}^{N}$ and $C_A = \{C_{a_i}\}_{i=1}^{N}$.

3.1.2 CoT Prediction

In this study, we consider the CoT prediction task as a sequential decision-making process and a reinforcement learning-based approach is empolyed to predict the logical structure, i.e., the CoT. Given the fact-rule CoT 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:

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

Then, we utilize a policy network $\pi_\theta(n_t|s_t)$ to generate the CoT 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 CoT node) predicted by the policy network. The initial CoT $C_{q_i}^{t=0}$ 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 CoT:

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

Subsequently, the state embedding is updated using the new CoT:

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

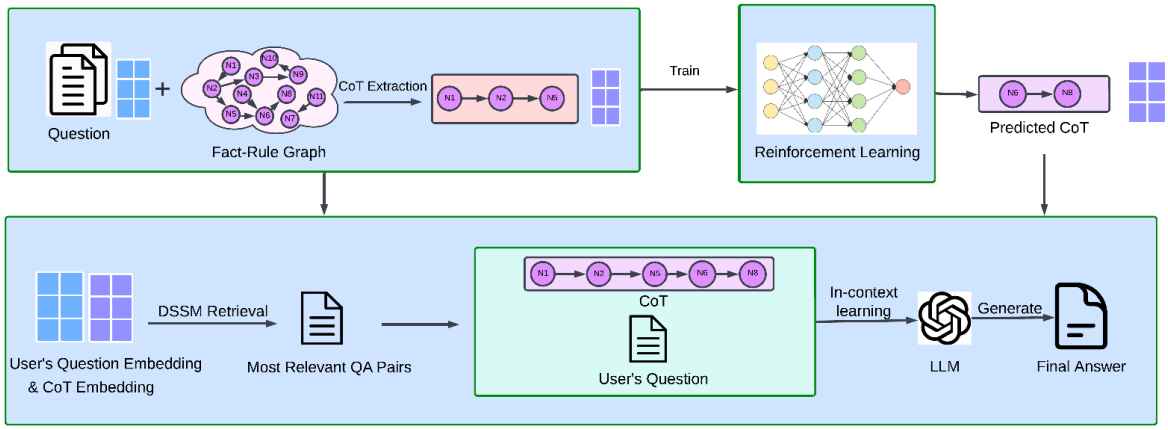

Figure 1: The overall framework of LSIM.

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 training objective is to maximize the expected cumulative reward:

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

(θ and π should be mentioned) 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 CoT for answer a_i.

During inference, the trained policy network is employed to generate the CoT for a given legal question. Assume the inference step is z, the generated CoT is $C^z_{q_i}$ for question q_i, and $C^z_{q_i}$ is the predicted logical structure.

3.2 Retrieval Model

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 $D = \{(q^D_j, a^D_j)\}_{j=1}^{M}$ 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 1 to 5, where a score of 1 indicates minimal relevance and a score of 5 denotes the highest level of similarity. 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 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.

Then, the current question q_i, the logical structure C_{q_i}, and $context_i$ are provided to the LLM to generate an answer:

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

4 Experiments

4.1 Datasets

We use real-world legal question and answer (Q&A) data and it was collected from JUSTIA[2]. The data comprises 16,190 legal questions posed by users in the area of criminal law, and each question has responses from at least one lawyer. The data is divided into database, training, testing sets. The specific information is presented in Table 1.

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

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)** and **Roberta (Liu et al., 2019)**, both pre-trained language models, are employed to generate embeddings for the user's query and all questions in the database. Similarity calculations are then used to determine the closest match. The classic RAG algorithms (Lewis et al., 2020) are also compared. Three advanced embeddings are used, i.e., text-embedding-ada-002, text-embedding-3-small, and text-embedding-3-large. **text-embedding-ada-002**, a more advanced embedding model developed by OpenAI[3]. **text-embedding-3-small** and **text-embedding-3-large**, the newest and highest performing embedding models developed by OpenAI. They are used to generate high-quality embeddings for the user's questions and the questions in the database. **LLaMA-2-13B** and **LLaMA-3-8B**[4], developed by Meta, 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 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

[3]https://platform.openai.com/docs/guides/embeddings/
[4]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
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
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 (%).

Method	Acc.	Spec.	Adopt.
text-embedding-3-large	4.35	4.35	4.41
LSIM	4.65	4.47	4.65

Table 4: Results of human evaluation

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 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. The experiments are conducted on an NVIDIA A100 GPU.

4.4 Experiment Results

Tables 2 and 3 present the experimental results obtained when Llama2-13B and Llama3-8B 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 performance 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-

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

User's Question
Someone I know has been accused of touching a child.Shes 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 shes 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?
Lawyer's Answer
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.
Response given by LLaMA-2-13B w/o RAG
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.
Response given by LSIM
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.

Table 6: Case Study

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. Compared to the best performing baseline model text-embedding-3-large, LSIM model achieves improvements of 2.52%, 1.97%, 0.35%, 1.59%, and 1.50% on METEOR, ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore.

The results of the human evaluation are shown in Table 4. Our LSIM model achieves the best performance in terms of accuracy, specificity, and adoptability, with scores of4.65, 4.47, and 4.65, respectively. The text-embedding-3-large model achieves 4.35, 4.35, and 4.41 points in accuracy, specificity, and adoptability, respectively. Compared to the text-embedding-3-large model, the LSIM model achieves improvements of 0.30, 0.12, and 0.24 in accuracy, specificity, and adoptability, respectively. These results highlight the effectiveness of our proposed LSIM model. The LSIM model is capable of retrieving the most relevant candidate questions for the given queries and generating highly pertinent and practical responses.

4.5 Ablation Study

An ablation study is also conducted for LSIM ton when using LLaMA-3-8B as baseline LLM. 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

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 7: The most relevant questions retrieved by LSIM.

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.

4.6 Case Study

Table 6 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. Table 7 presents the most relevant question retrieved by LSIM. The answer to this question advises the user to "keep your mouth shut" an "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.

5 Conclusion

In this paper, we address the limitations of LLMs in providing expert-level responses to legal questions. We propose an novel framework named LSIM, which integrates logical structure and semantic information of the legal question to enhance the ability of LLMs to generate expert legal responses. The LSIM framework consists of three components. First, reinforcement learning is used to predict the fact-rule chain of thought for the given question. Second, the DSSM model that integrates logical structure and semantic information is used to retrieve the most relevant candidate questions from the database. Finally, the chain of thought for the user's question, the most relevant retrieved questions, and their corresponding answers are provided as reference information to the LLM to generate the final answer. Experiments are conducted on a real dataset. Both automated evaluation metrics and manual evaluation demonstrate the effectiveness of our proposed LSIM framework.

In the future, we will explore applying our model to other specific domains, such as medicine and finance. Additionally, we will explore adopting multiple turns of interaction with users and incorporating their real-time feedback into the training process of our model to further enhance the performance.

Limitations

The limitations of our work are as follows:

- 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.
- Our study is limited to single-turn interactions with large language models. Multi-turn interactions could potentially help the model gain a more comprehensive understanding of the questions and provide more accurate and targeted answers.

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.

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. Mlljp: 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.

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*.

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.