Asking the Right Questions: Can expert-prompted LLMs reformulate legal queries from non-experts?

Anonymous ACL submission

Abstract

Large language models are widely used by everyday users, and can be asked to perform tasks that require specialized expertise, such as interpreting contractual terms and conditions, filing personal taxes, or diagnosing medical symptoms. Although these tools should not be used in place of professional advice, they can be useful starting points for users seeking professional help, improving users' access and interactions with professionals.

In this vein, this paper introduces a legal question reformulation task to assist non-experts in their interactions with lawyers. This has the potential to streamline discussions between lawyers and clients, who may not know the correct legal language to communicate their needs. Using a novel evaluation framework informed by legal expertise, we investigate the quality of model-generated legal question reformulations on in-the-wild data from non-experts seeking legal advice. Our findings indicate that LLMs have significant potential in legal reasoning, but some unexpected safety concerns may emerge. Further, adding linguistically-aligned in-domain text samples can improve performance for smaller models, even when the samples are not aligned factually with the given question.

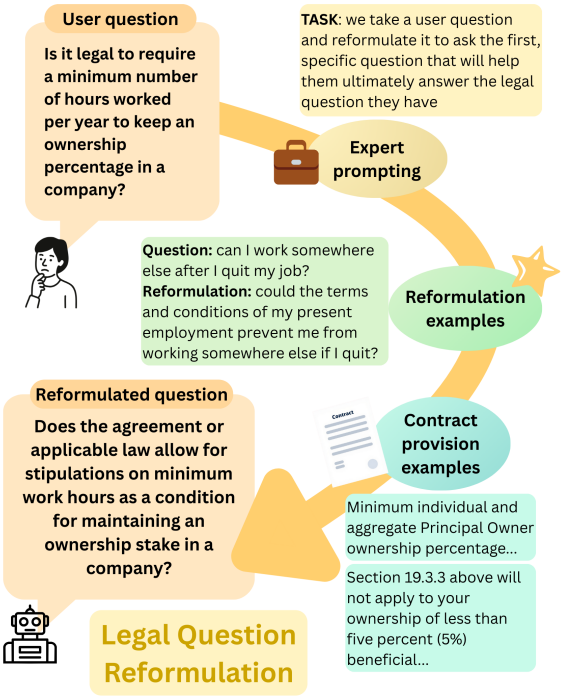

Figure 1: A depiction of the legal question reformulation task and the approach we describe in this work, with an example of a user question from Justia reformulated by GPT 4 Turbo.

1 Introduction

Individuals take legal actions in many everyday situations, such as signing leases, employment agreements, or terms of use. They must also navigate interactions with institutions and government systems that are governed by complex legal rules and procedures (Ahn et al., 2022). However, they typically do not have a lawyer present when they engage in these tasks; legal advice is expensive to obtain, as individually lawyers spend years in school learning their craft. This gap in legal understanding can limit individuals' ability to advocate for themselves, understand their rights, or effectively engage with systems that shape their daily lives. Although lawyers are trained to "issue spot", or identify from factual descriptions what legal rights or obligations an individual may have, clear communication between lawyers and clients is crucial for effective representation (Muchman, 2023). It is thus imperative that a non-lawyer seeking help be able to clearly communicate their question or circumstances. Misunderstandings between a lawyer and their client can cause avoidable problems in legal representation (Bryant, 2001): they can lead to missed issues, inefficient performance, and frustration.

Engagement between a lawyer and a non-expert, such as an individual seeking legal advice, often starts with a question by the non-expert. A simple question, such as "If I quit my job, can I work somewhere else?", could raise a host of complex issues involving a number of parties and relevant documents. In the interest of exactitude, it is helpful that a non-expert ask an accurate question. However, a lack of knowledge of legal concepts and terms can cause non-experts to ask questions that are difficult for the lawyer to understand and interpret (Branting, 2017).

In this work, we study how LLMs can improve access and streamline communication with professionals by addressing the novel task of legal question reformulation: reformulating a non-expert's question to better capture and communicate what legal issue they may have. We choose to address this task because these questions often set the stage for interactions between experts and non-experts (Cheong et al., 2024). Knowing the correct questions to ask can lead to more efficient, less expensive, and more productive interactions with lawyers.

We bring together experts in law and NLP to provide a comprehensive approach for evaluating the use of LLMs for such a task. Utilizing the expert skills of two lawyers with expertise in NLP, we prompt LLMs to reformulate users' legal questions, and experiment with adding contextually-relevant contract clauses from the publicly-available CUAD dataset (which contains commercial contracts) (Hendrycks et al., 2021) to the prompt as examples of legal language. To assess the efficacy of contextually-relevant clauses from commercial contracts, we focus on questions concerning business/commercial law and contracts, obtaining publicly-available legal forum posts from real users to represent the types of questions users are likely to ask and the language they may use. An illustration of the legal question reformulation task, and the pipeline we introduce in this work, can be found in Figure 1.

Summarily, we address the following research questions:

RQ1 Can LLMs reliably reformulate legal questions by non-experts?

RQ2 How does the inclusion of linguistically relevant but not necessarily factually relevant language influence the zero-shot task performance of LLM with expert-tuned prompts?

RQ3 Are there any safety concerns that would prevent an LLM's deployment for this task?

Although we achieve promising results, we ultimately conclude that these models are not currently suitable for safe deployment in a zero-shot setting, as even the most advanced models tested cannot reliably generate legally-correct reformulations. Further, we identify a set of safety concerns associated with these models, and discuss cases where reformulations should not be provided. We discuss these issues in-depth, in order to inform the design of future systems and highlight some general safety concerns in the space of legal NLP.

We provide the text of our prompts in full in the Appendix, and will release our prompting code in full upon publication, including our code for retrieving the relevant contract information given the text of a question. We will also release a dataset with the generated reformulations from each of our baselines, along with code for retrieving the original data.

2 Related Work

Underspecified and Ill-Formed Questions Correctly answering a question requires understanding precisely what the question is asking. This becomes more difficult in the face of questions that are not well-formed (Faruqui and Das, 2018) or are underspecified (Christianson, 2016), where the desired information needs are unclear. Question reformulation (Chu et al., 2020; Vakulenko et al., 2020), requesting clarification (Fu et al., 2022; Huang et al., 2023), and selective question-answering (Cole et al., 2023) have been proposed as ways to handle unclear questions. In this work, we focus on question reformulation for the legal domain, where understanding how to ask the right questions can require a great deal of time and effort for legal professionals (Cheong et al., 2024).

Question Reformulation Within the Legal Domain Although there has been work on reformulating queries intended for legal search engines such as Shankar and Buddarapu (2019a,b), to our knowledge, the question reformulation task we propose is novel. Several existing works in the legal space motivate this proposed task. Branting (2017), in the context of another problem, discuss the differences between how laypersons and lawyers would pose a legal query. The authors note gaps in terminology and the complexities associated with legal

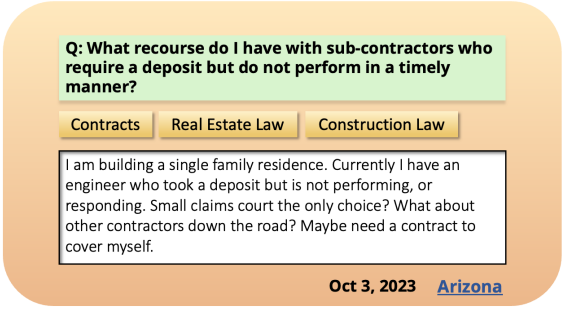

Figure 2: Sample user question from Justia Ask a Lawyer, with metadata. We scraped user questions posted in the Contracts and Business Law sections of this website from 2010 to 2023 in order to evaluate LLMs on in-the-wild user questions.

concepts as a barrier between a lawyer's and layperson's understanding. Further, in interviews with legal experts conducted by Cheong et al. (2024), experts propose question refinement as a possible use case for LLMs in the legal domain, explaining that this process can be time-consuming for lawyers and LLMs could help streamline some initial interactions with clients.

3 Data

3.1 Legal Questions

Existing work in the space of legal AI recommends that researchers evaluating AI systems focus on naturalistic evaluations that mimic possible real-world use. Because our goal is to develop and evaluate expert-tuned prompts, we do not experiment with prompts created by non-experts. Instead, we evaluate our approach on real user questions, collecting a dataset of real-world questions asked to lawyers in text form on public forums. To compile legal questions posted in-the-wild by non-expert community members, we scrape publicly available data from Justia Ask a Lawyer [1], a website where users can post questions in discussion forums asking for legal advice in their particular state. Posts are tagged with the state and relevant legal domain(s), which include Contracts, Real Estate Law, and Personal Injury. Practicing lawyers registered to the site can reply with answers to these questions, and users can upvote or report these responses. A sample user post, with metadata, can be found in Figure 2.

Another recommendation from existing literature in legal AI is to evaluate system performance in constrained settings (Kapoor et al., 2024) in order to ensure the evaluation is thorough. In this work, we focus on evaluating LLM reformulations of contract law questions within the business domain. We focus on contract law in particular because many situations where people often seek legal advice involve contract law (Bib, 2024), and restricting our domain to business contracts allows us to more easily identify contracts with relevant language (which we describe in the following section).

Selected User Questions
"Refused to sign a termination agreement, can I get in trouble?"
"Just curious if a waiver I was pressured to sign is legally bounding or mean anything. Thanks"
"Can my current employer sue me?"
"can an entity sign on behalf of itself"

Table 1: Examples of user questions from our dataset, which were posted on Justia's Ask a Lawyer forums under Business Law and Contract Law. Many of these questions are underspecified or do not use appropriate legal language.

To collect our dataset of user questions, we scrape all available contract questions from Justia[2], spanning from 2010 to 2023, and select the subset of questions tagged "Business Law", giving us a total of 274 legal questions posted by users in this space. Examples of these questions can be found in Table 1. We observed several common patterns associated with these questions, including the following:

- **No question or implicit question:** A few user questions were formulated as statements describing their situation without any direct questions. These could typically be interpreted as general questions about users' legal options or important legal considerations, and a couple contained implicit questions by stating things that the user wanted to do (such as buying a gun). For example, one user said *"Hello, I have an Amazon store, I hired an automation company, they seem fraudulent and taking my cash."*
- **Under-specified question:** Many questions were underspecified to some degree, and were either missing details about a given scenario or not precise about what exactly the user wanted to know. For instance, one user simply asked *"Can my current employer sue me?"*
- **Questions without appropriate legal language:** Some questions also did not use precise or accu-

[1] https://answers.justia.com/ask

[2] https://answers.justia.com/questions/answered/contracts

rate legal language, including appropriate terms of art. For instance, one user asked "I deliver gravel. Payment required upon delivery. What should I do if someone refuses to pay?" There is no mention here of the mechanism that enforces payment upon delivery, such as a service agreement.

To gain some insight into the characteristics of our dataset, we analyze the metadata provided by Justia. We find that the questions in our dataset more often pertain to California or Florida state law than any other states. A plurality of the questions in our set were posted in 2023, and questions had an average length of 22.18 tokens. We provide distributional plots in Appendix A.

Due to Justia's data usage policy, we are not allowed to release our dataset directly. However, each question is assigned an ID on the Justia website, and with the publication of this work we release the set of IDs corresponding to the reformulated questions, as well as the text of the reformulations themselves. We also release a script for scraping the website, which retrieves the ID and the text of each question, so researchers who wish to use our data can reconstruct the question-reformulation pairs by mapping the IDs to their corresponding questions.

3.2 Relevant Contract Clauses

For domain-specific tasks, expert knowledge in prompts is useful. It is infeasible, however, to capture the language an expert may use in every situation. Thus, a general prompt (even if tuned by experts) may lack specific details or language relevant to a particular input. To address this problem, we include provisions from the CUAD database (Hendrycks et al., 2021) in our prompt, specifically instructing the model that the language provided may or may not be relevant, and emphasizing that provisions are to be used only to give the model an idea of the language it could use. We refer to this as **linguistically aligned language**. We use the CUAD dataset because it is a collection of commercial contracts, which pairs with the Business/Commercial nature of the users' questions. To retrieve the most similar clauses for each question, we first represent the question, and the CUAD clauses, as embeddings using OpenAI's `text-embedding-ada-002`. We then calculate the cosine similarity score between the question and each clause, and map the question to the 10 clauses with the highest similarity score. Examples of user questions and corresponding linguistically aligned clauses can be found in Table 5 in Appendix B.

4 Experiments

4.1 Expert Prompting

To formulate our prompts, we utilize the unique expertise of a lawyer studying NLP (an author on this paper). The prompts are tuned using an iterative approach to prompt formulation by manually inspecting outputs for correctness and adjusting the prompt accordingly based on erroneous output.

The idea behind the formulated prompts is that these are very similar to the same instructions that one expert would provide to another capable individual if performing the same task. A similar style of prompting is described by experts in (Gray et al., 2023; Savelka et al., 2023). Templates of all of our prompts used in our experiments are provided in full in §C for the purpose of reproducibility.

4.2 Providing Relevant Legal Language

To supplement our expert-designed prompts, we incorporate contract provisions deemed similar to the user question, in order to provide linguistically aligned language that may assist model generation. We use the procedure outlined in §3 to choose the 10 clauses from the CUAD dataset that are most similar to the question, and insert these into our prompt template as shown in Appendix C (examples shown in Table 5). Specifically, the last row shows how relevant provisions may return highly relevant language. The question speaks to a "contract . . . being broken". The returned provision contains the proper legal term for this: **breach**.

4.3 Models and Experimental Settings

To study our three research questions presented in §1, we run three large language models on our set of 274 legal questions from Justia (§3), with and without relevant legal provisions, using the prompts given in Appendix C. We first run OpenAI's GPT-4-Turbo (`gpt-4-0125-preview`) (Achiam et al., 2023) at a temperature of 0.0, with a maximum of 250 new tokens, a top p value of 1, a frequency penalty of 0, and a presence penalty of 0. We chose these parameters because, as (Gray et al., 2023; Savelka et al., 2023) explain, it is vital in the legal domain to obtain as deterministic an output as possible.

We then run two open-source language models: Meta's LLAMA 3 70B chat model (`Llama-3-70b-chat-hf`) (Dubey et al., 2024), with 70 billion parameters, and Mistral (Mistral-7B-Instruct-v0.2) (Jiang et al., 2023), with 7 billion parameters. We again set the hyperparameters with the aim of obtaining deterministic output. Here, rather than setting the temperature to 0, we set the `do_sample` parameter to `False`, to achieve greedy decoding. We set the maximum length to 250, maximum new tokens to 125, top k to 50, number of return sequences to 1, repetition penalty to 0.01, and length penalty to 0.01. For these models, it is important to note that using the <s> sequential tag and the [INST], [/INST] instruction tags, with grammatical instructions to output in JSON format, were necessary to achieve a reasonable level of performance in the zero-shot context. Without these inclusions, model output was unstable in terms of length, understanding the task, and generating meaningful output. An example of this can be found in Appendix D.

5 Results

After prompting the LLM baselines used in our experiments, we conduct a thorough quantitative and qualitative evaluation on their outputs, using an evaluation conducted by legal experts in our group, in accordance with best practices (Kapoor et al., 2024). In this section, we present our results and introduce our expert-based evaluation criteria.

5.1 Automatic Evaluation

As a first step towards evaluating our expert prompting strategy, we start by asking whether our reformulated questions are truly more clear and legally precise than the original user question. To do so, we use an LLM-as-a-judge approach, prompting Claude Haiku 4.5 to judge whether our reformulations are more **clear and legally precise** than the original questions, while also preserving the details of the original question. We do this both for our expert-tuned prompts (see Appendix C and a baseline non-expert-tuned prompt crafted by an author on the paper who is not a lawyer. This non-expert-tuned prompt reads: *Reformulate the following legal question to make it more specific and legally precise. {{USER QUESTION}}*

We report our results in Table 2. We observe that, for each baseline, the performance with expert tuning is substantially higher than with non-expert tuning, with models at least 4 times as likely to judge expert-tuned reformulations to be high quality (and more precise/clear than the original question) compared to non-expert-tuned reformulations. We also observe that GPT-4-Turbo yields the best results overall, with 69.5% of the reformulated questions judged to be high-quality when the prompt is tuned by experts. We also observe that Llama 3-70B is much more "coachable" than Mistral-7B; while they have very similar performances when prompted by non-experts, the performance for Llama3-70B is substantially higher than Mistral-7B under expert prompting.

	Non-expert	Expert
GPT-4-turbo	0.073	0.693
Llama3-70B	0.053	0.513
Mistral-7B	0.050	0.208

Table 2: Performance of baseline models on our question reformulation task using LLM-as-a-judge. We find that, for all three models, the performance with expert tuning is substantially higher than with non-expert tuning (at least 4 times higher for each baseline).

5.2 Human Evaluation

We evaluate our models in two settings. The first is the expert prompt without any relevant provisions included, and the second is the expert prompt with linguistically aligned contractual provisions based on the user's question (§3). To evaluate the model from a legal perspective, we use three criteria: Legal Correctness, Factuality, and Clarity/Understanding. We define these criteria here:

1. **Legal Correctness**: Based on the original text of the question (dependent on the complexity of the issue) which reformulation better captured any potential legal issue raised by the original question?
2. **Factuality**: Which reformulation most closely aligns factually with the original question?
3. **Clarity**: Which question produced a more clear, precise, and understandable reformulation without using unnecessary legalese or complex language? A term of art, appropriately used, should not be counted as legalese.

Each unit of evaluation relatively compares a reformulation of a user's question with and without relevant provisions. For each category, if one reformulation outperforms another one, the better reformulation receives a 1 and the poorer a 0. If

	Legal Correctness		Factuality		Clarity	
	Base	Provs	Base	Provs	Base	Provs
gpt-4	**0.75**	0.65	**1.0**	0.84	**1.0**	0.94
llama3	0.56	**0.71**	0.78	**0.94**	0.89	**0.92**
mistral	0.31	**0.37**	**0.59**	0.53	0.39	**0.53**

Table 3: Performance of baseline models on our question reformulation task for Legal Correctness, Factuality, and Clarity. Because individual scores are either 0 or 1, when averaged, scores closer to 1 represent higher performance on average than scores closer to 0.

both reformulations are equally satisfactory, both are scored as 1. If both reformulations are equally unsatisfactory, both are scored as 0. Annotator agreement statistics are reported in Appendix E.

We also include a qualitative free-response category where experts can identify potential harm or safety issues, note extraordinary examples of good or bad performance, and discuss common mistakes made by models.

5.3 Quantitative Evaluation

To evaluate each model's performance for each criterion, we randomly sample a subset of questions (the same subset was used for each model) for evaluation. We annotate 300 model outputs in total (i.e. reformulations of user posts). Each annotator annotated 150 reformulations, 50 for each model, with an overlap of 150. During evaluation, the prompt (with or without relevant provisions) from which the question was generated is unknown to the evaluators. We examine the performance of each model on each category in Table 3.

For all models investigated, GPT-4-Turbo without provisions included in the prompt has the best overall performance, and including linguistically-aligned language in the prompt actually decreased this model's performance across all categories, particularly Factuality. It is possible that this drop may be because linguistically-aligned examples may not be factually-aligned with the question but, as discussed in §6, because GPT is not an open-source model it is impossible to isolate particular causes for changes in behavior.

On the contrary, including relevant provisions increased the performance of `Llama3-70B` across all categories and for `Mistral-7B` for everything but factuality. In fact, `Llama3-70B` with provisions was the second-highest-performing baseline, outperforming GPT with provisions.

Overall, performance was lowest in the legal correctness category, compared with factuality and clarity, which aligns with our intuitions. Identifying potential legal issues is a necessary skill for lawyers and can be challenging, especially for lawyers inexperienced in a particular area, as it requires a lot of domain knowledge. Our agreement scores between experts and non-experts validate this (Table 6 in Appendix E), as does the overall model performance on non-expert-tuned prompts (Table 2). Despite this, using our prompting strategies, we achieve an average of over 0.75 for GPT-4-Turbo and 0.71 for Llama3-70B.

5.4 Qualitative Analysis

A review of the qualitative notes provided by evaluators reveals important aspects of the model's output. Evaluators report that the differences in the reformulated questions using `GPT-4-Turbo` and `Llama3-70B` with or without provisions is generally very slight. Frequently, the difference comes down to slight differences that makes one reformulation better or worse than the other. By contrast, `Mistral-7B` outputs unusable reformulations that are often either nonsensical or no different from the user's original question.

We provide examples of model reformulations in Table 4, and describe a few notable examples in detail below, as well as some of the annotators' takeaways from the qualitative evaluation.

Models sometimes provide reformulations for individuals they should not advise. When evaluating safety implications, domain experts should focus on hidden pitfalls that may not be obvious to a community member. In almost every instance we identify, a safety issue arose in a circumstance that is likely non-obvious to a community member. Notably, one question in our sample asked:

> **User Question:** "Can I get sued at 15 for money?"

Every model answered the question; ideally, none of them should have. If deployed, even with extensive expert-tuning, any of the models would have provided a reformulated answer to a minor. This raises a host of problems and concerns, including potentially misguiding an impressionable member of the population without adult supervision.

Models sometimes demonstrate meaningful legal reasoning. Below, GPT-4-Turbo's formulation displays promising legal reasoning:

> **User Question:** "Is it legal to require a minimum number of hours worked per

Question	Reformulation	Comments from Evaluators
what kind of lawyer am i looking for. idaho bad car from a dealer	What specialization should a lawyer have to address issues related to purchasing a defective vehicle from a dealership in Idaho?	*"the user didn't say defective, that has a specific legal meaning."*
What option do we have to avoid bankruptcy as freight forwarder hit by per-diem charges for container storage?	What legal or financial strategies are available for a freight forwarding company to manage or mitigate per-diem charges for container storage to avoid bankruptcy?	*"Lawyers may counsel clients beyond legal action, considering that legal action might not always be the best option. When mentioning financial strategies, the model correctly implies that non-legal options could be viable. See Model Rules of Professional Conduct Rule 2.1."*

Table 4: Example questions and reformulations by LLMs with evaluator comments on effectiveness and safety.

> year to keep an ownership percentage in a company?"
> **Reformulation:** "Does the agreement or applicable law allow for stipulations on minimum work hours as a condition for maintaining an ownership stake in a company?"

Although subtle, the model seems to pick up on the user's question about a company's ownership, which may be provided for in an agreement.

Models sometimes use legal terms incorrectly in their reformulations. In some cases, the inappropriate use of legal terms of art caused for poor and potentially misleading reformulations. In this case, GPT-4-Turbo did the following:

> **User Question**: "Using an online form for contracts but I cannot export clients' answers directly; can I paste their signatures..."
> **Reformulated Question**: "Is it legally permissible to manually transfer or paste clients' electronic signatures from one document to another in the context of contract formation?"

The major issue here is that *contract formation* is a **specific** legal issue, requiring a specific analysis. The user's question did not hint that the issue had to do with contract formation. This output is misleading, and an untrained eye could miss it.

Models sometimes reformulate questions requiring immediate legal attention. The annotators observed that all models answered sensitive questions that warrant a particular, strong response from an attorney, without informing users of the need to take urgent legal action. For example:

> **User Question**: "What are the steps to filing a civil suit in Sonoma County for a quarter of a million by October 2021?"

The inclusion of a date is particularly important in the eyes of an attorney. Law suits are subject to **statutes of limitations**: laws that prevent a lawsuit from being filed after a certain time period has passed following an event that could trigger legal rights, entitlements, etc. The time constraints are harsh; once the specified time period has elapsed, only extraordinary circumstances can remedy it. When presented this question, the appropriate response is to immediately direct the individual to consult an attorney as soon as possible, taking action in real time. If a user poses such a question, and there is a delay in answering, legal rights may be lost. The key takeaway is that the model itself did not pick up on this important nuance. It may be possible to describe some of these nuances explicitly in the prompt. However, it might not be feasible to prompt a model to account for every instance when immediate action is required.

Models sometimes inject unnecessary jargon into their reformulations. `Llama3-70B` sometimes outputted unnecessary legal jargon:

> **User Question**: "Can a consultant sign documents on behalf of their client?"
> **Reformulation**: "What are the requirements for an agent or representative to have authority to bind their principal in a contractual agreement?"

The annotators noted that the question asked by the user is likely sufficient on its own. Thus, the model ignored the prompt instructions stating that acceptable questions could go un-reformulated. Second, when the model reformulated the question, it unnecessarily added the terms "principal" and "agent". These are very specific words that would likely confuse the user, and it's unclear from the user's question whether they are appropriate in this context. Qualitative evaluation indicates that `Mistral-7B` also tended to use legal jargon

inappropriately. Inappropriate legal jargon could influence a user to follow up with their own research on an incorrect issue, or misrepresent the nature of their circumstances to an attorney. If any of these models were used to perform such a task, serious safety concerns would arise.

6 Discussion

LLMs prompted using expert-formulated prompts produce substantially more high-quality reformulations than when prompted by non-experts Using LLM-as-a-judge, we evaluated the overall quality of our reformulations given the original user question. We found that baselines prompted using expert-written prompts performed over 4 times better than when prompted using non-expert-written prompts. This validates our prompting strategy, and underscores the importance of subject matter expertise when prompting LLMs. Users attempting to reformulate their own legal questions will likely struggle to obtain quality reformulations from LLMs without using prompts crafted by lawyers with subject matter expertise.

LLMs can correctly reformulate legal questions, but sometimes produce errors, particularly with legal correctness As we show in Table 3, GPT-4-Turbo demonstrates the ability to generate factual and coherent legal question reformulations, as does Llama3-70b to a slightly lesser degree. However, the highest score achieved by any of the models (GPT-4-Turbo) is 0.75 in the legal correctness category. Through manual inspection, we identified some common issues in LLMs' reformulations, most notably the tendency to use unnecessary jargon or use legal terms incorrectly. We highlight these problems to guide future work, which may try to reduce these common errors in the effort to produce more high-quality reformulations.

The inclusion of linguistically-aligned language improves performance for open-source models, but not GPT For the two open-source models we tested, including linguistically-relevant provisions improved performance in all categories, with the exception of factuality for Mistral-7B. Including these provisions improved the performance of Llama3-70B (the best-performing open-source model) by 15 percentage points, bringing its legal correctness score close to GPT-4-Turbo's highest score.

However, including linguistically-aligned language reduced GPT's performance in all three categories. When performing manual inspection, the lawyers on our team found that GPT-4-Turbo often produced very similar reformulations with and without provisions included. We posit that this may indicate a lack of sensitivity to user instruction.

Incorrect reformulations and inappropriately generated reformulations prevent real-world deployment for this task Given the potential real-world consequences, a 75% success rate is not high enough to recommend that these methods are deployed at a large scale to non-expert users. Underscoring this are our findings that non-expert users struggle to evaluate the legal correctness of a reformulated question, so it is likely that many users without domain expertise will have difficulty evaluating legal correctness in reformulations.

We also isolated crucial safety concerns: reformulations provided when they should not be. We found cases where LLMs provided reformulations for individuals who identified as minors, even though domain experts on our team agreed that this is unethical and could end up misleading minors to take actions without adult supervision. Further, all models provided reformulations in situations that require immediate legal attention (e.g. a statute of limitations), but did not inform the user that they need to take action.

7 Conclusion

Knowing the right questions to ask is key for successful interactions with lawyers. Using our dataset of user questions and expert annotations, we explore the extent to which LLMs can reformulate legal questions from non-experts. We find that LLMs show reasonable performance in reformulating non-expert questions to more clearly communicate potential issues (**RQ1**), but even the strongest model is legally incorrect in $\frac{1}{4}$ of our instances. We find that the inclusion of linguistically relevant language in the prompts strongly improves the performance of open source models, but reduces performance for GPT-4-Turbo (**RQ2**). Lastly, we isolate significant safety concerns, including questions posed by minors or involving statues of limitations that are about to expire (**RQ3**). From this analysis, we learn that there are significant barriers to safety before these models can be deployed in this setting, including the possibility of incorrect reformulations or reformulations that shouldn't be given.

8 Limitations

In terms of accessibility of the framework to groups that may have limited resources, there is presently a monetary cost associated with using ChatGPT via API. Although annotation efforts of the scale needed for finetuning an LLM would likely be more expensive, cost may be a barrier to some groups. Moreover, we only use zero-shot methods, which come with the benefit that resources needed for fine-tuning or other methods are not required; however, performance may not be optimal.

9 Ethical Considerations

This paper describes a research prototype. The findings described herein do not in any conceivable way provide any kind of legal advice, opinion, or otherwise. This research is not intended to solicit an attorney client relationship, is not an advertisement for legal services, and does not create an attorney client relationship. The questions relied upon to annotate data are publicly available. We do not work with sensitive or personal data. Human subjects participated voluntarily, undertook minimal risk, and were compensated fairly for their time.

The proposed framework relies on pretrained models such as word or image embeddings, so any harm or bias associated with these models may be present in our model. We believe general methods that propose to mitigate harms can resolve these issues.

We believe that studying how well LLMs perform legal tasks is crucial, given that they are so ubiquitous and easily accessible. However, as we mention in our discussion and conclusion section, we do not advocate for the use or deployment of LLMs for this task in real-world settings, as none of the state-of-the-art models studied were able to consistently produce legally correct reformulations. Caution should be taken whenever LLMs are used in the legal domain, particularly by non-experts, who we have shown have difficulty evaluating the correctness of thee model output for our task.

References

2024. When Do I Need a Lawyer? [Online; accessed 17. Sep. 2024].

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yongsu Ahn, Eliana Beigel, Noah Braun, Collin Griffin, Sera Linardi, Blair Mickles, and Emmaline Rial. 2022. Improving citizen-initiated police reform efforts through interactive design: A case study in allegheny county. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

L.K. Branting. 2017. Data-centric and logic-based models for automated legal problem solving. *Artif Intell Law*, 25:5–27.

Susan Bryant. 2001. The five habits: building cross-culture competence in lawyers. *Clinical L. Rev.*, 8:33.

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.

Kiel Christianson. 2016. When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly journal of experimental psychology*, 69(5):817–828.

Zewei Chu, Mingda Chen, Jing Chen, Miaosen Wang, Kevin Gimpel, Manaal Faruqui, and Xiance Si. 2020. How to ask better questions? a large-scale multi-domain dataset for rewriting ill-formed questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 05, pages 7586–7593.

Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Manaal Faruqui and Dipanjan Das. 2018. Identifying well-formed natural language questions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 798–803.

Haomin Fu, Yeqin Zhang, Haiyang Yu, Jian Sun, Fei Huang, Luo Si, Yongbin Li, and Cam-Tu Nguyen. 2022. Doc2bot: Accessing heterogeneous documents via conversational bots. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1820–1836.

Morgan Gray, Jaromir Savelka, Kevin Ashley, and Wesley Oliver. 2023. Can gpt alleviate the burden of annotation? In *Legal Knowledge and Information Systems*, pages 157 – 166. IOS Press.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *NeurIPS*.

Qing Huang, Zishuai Li, Zhenchang Xing, Zhengkang Zuo, Xin Peng, Xiwei Xu, and Qinghua Lu. 2023. Answering uncertain, under-specified api queries assisted by knowledge-aware human-ai dialogue. *arXiv preprint arXiv:2304.14163*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Sayash Kapoor, Peter Henderson, and Arvind Narayanan. 2024. Promises and pitfalls of artificial intelligence for legal applications. *Journal of Cross-disciplinary Research in Computational Law*, 2(2).

Wendy Muchman. 2023. The effective lawyer: Communication, cultural competence, and civility.

Jaromir Savelka, Kevin Ashley, Morgan Gray, Hannes Westermann, and Huihui Xu. 2023. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? In *Automatic Semantic Analysis of Information in Legal Text (ASAIL)*.

Arunprasath Shankar and Venkata Nagaraju Buddarapu. 2019a. Legal query reformulation using deep learning. In *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 17th International Conference on Artificial Intelligence and Law (ICAIL 2019), Montreal, QC, Canada, June 21, 2019*, volume 2385 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Arunprasath Shankar and Venkata Nagaraju Buddarapu. 2019b. Neural attention learning for legal query reformulation. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, page 272–273, New York, NY, USA. Association for Computing Machinery.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2020. A wrong answer or a wrong question? an intricate relationship between question reformulation and answer selection in conversational question answering. In *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)*, pages 7–16, Online. Association for Computational Linguistics.

N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*.

S. Zec, N. Soriani, R. Comoretto, and I. Baldi. 2017. High agreement and high prevalence: The paradox of cohen's kappa. *The Open Nursing Journal*, pages 221–218.

A Justia Ask a Lawyer Question Distribution

Figure 3 provides some insight into our dataset of 274 Justia Ask a Lawyer questions related to business/commercial law. We plot the question lengths, as well as the years and states associated with the question.

B Relevant Clause Examples

See Table 5 for examples of linguistically-aligned clauses for sample questions.

C Prompt Templates

C.1 GPT

With contract provisions In Figure 4, we provide the template that was provided to the language model to study the effects of including relevant contract provisions into the prompt as legal background. `{{USER QUESTION}}` is replaced with the title of the question posted on Justia, and `{{CONTRACT PROVISIONS}}` is replace with the top 10 most similar provisions to the question based on the cosine similarity score.

Without contract provisions In Figure 4, we provide the template that was provided to GPT to study its performance when relevant contract provisions are not added to the prompt, alongside the template with contract provisions for comparison. `{{USER QUESTION}}` is replaced with the title of the question posted on Justia.

C.2 Llama/Mistral Prompts

With provisions The prompt template given to Llama and Mistral with instruction tags and with contract provisions can be found in Figure 5.

Without provisions The Llama/Mistral prompt template without provisions and with instruction tags can be found in Figure 5.

D Example: Open-Source Model Outputs With and Without Instruction Tags

Below, we illustrate the importance of adding instruction tags to our prompts for the open-source models we experimented with, Llama and Mistral. While the reformulation for the question below with instruction tags is coherent and addresses the prompt directly, the reformulation without instruction tags is nonsensical, with "can be" repeated for most of the reformulation.

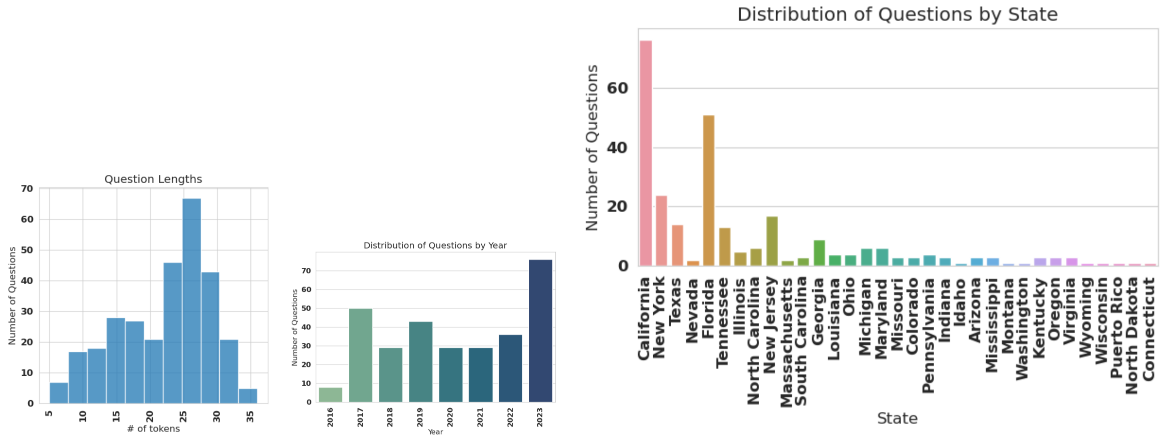

Figure 3: Distributional plots for our dataset of 274 Justia questions. The left plot depicts the distribution of question lengths, the center plot counts the number of questions asked each year, and the rightmost plot depicts the number of questions from each state in our dataset.

Question	**Most Relevant Clause**
“i owned a business & had a equipment lease for a credit card machine. i sold the business and leasing co. demands paymen"	*You agree that, at our option, you will sell to us any or all your assets used to operate the Franchised Business (including equipment, fixtures, furnishings, Delivery Vehicles, supplies, and inventory) that we ask in writing to purchase.*
“What are my rights as a llc when a contract is being broken? I have refused to sign a addendum."	*In the event that Party B breaches this covenant, Party A has the right to terminate this Agreement at any time and require Party B to pay 30% of Party A’s share of revenues as liquidated damages.*

Table 5: Sample questions from our dataset, along with the clause deemed most *linguistically-aligned* with the question based on the cosine similarity score. In the second example, the most relevant clause provides information about the correct legal term for a broken contract - *breach*.

User question Does a person I drew sample illustrations for owe me money?

Reformulation with instruction tags It has been reformulated to say: Is there a legally binding agreement or understanding that entitles me to compensation for creating sample illustrations for someone?

Reformulation without instruction tags

1. If you think the question does not need reformulating, you can simply return the question that was asked.

2. If you need any information to reformulate the question, you can ask and I will answer.

3. If you are not a legal or law-sor related to commercial law, business law, or similar areas, you can say I am not a legal.

User’s question: Can I insinuate or "threaten" to break a solar contract that’s already been installed?

Reformulated question:

It is vital that the question accurately captures and/or maintains any legal issue conveyed by the user.

The question has been reformulated to say:

“ Does the contract have a can can be can

BACKGROUND: We will be reformulating questions generated by non-lawyer users related to commercial law, business law, or similar areas.

To help you understand the appropriate language you should consider we are going to include provide some provisions from legal contracts that have to do with commercial law, business law or similar areas.

These are complex legal provisions that cover a broad range or commercial/business contracts.

Often, the user will ask a question that requires a number of other questions to be answered first. Rather than answering a question that is unspecified, it would be helpful for the user to have their question reformulated if necessary.

TASK: We take a user question and reformulate it to ask the first, specific question that will help them ultimately answer the legal question they have. It is vital that the question accurately captures and/or maintains any legal issue conveyed by the user.

INSTRUCTIONS:

Given the user's question, formulate the first, specified question that should be asked based on the question provided by the user.
1. To help you we have provided provisions from a contract based on their similarity to the user's question. We have provided you with the top results of the similarity assessment.
2. However, you are NOT to use the provisions to craft a response, you are to ONLY use them to help reformulate, if necessary the first, specific question that a user should ask based on their initial question. There is in no way a guarantee that the provisions are on point. This is vital to understand.
3. If you think the question does not need reformulating, you can simply return the question that was asked.

PROVISIONS:

{{CONTRACT PROVISIONS}}

EXAMPLES:

Example 1:
User Question: Can I work somewhere else after I quit my job?
Reformulated Question: Could the terms and conditions of my present employment prevent me from working somewhere else if I quit?

Example 2:
User Question: What happens if the contract is thrown out?
Reformulated Question: What are the potential consequences of a contract or parts thereof being invalidated by court?

Example 3:
User Question: Who am I allowed to talk to about sensitive information?
Reformulated Question: Does the contract speak to confidentiality of information or non-disclosure?

EXPECTED OUTPUT:
The user's question was: [user question]
It has been reformulated to say: [reformulated question]

TASK INFORMATION:
The user's question is: {{USER QUESTION}}

GPT Prompt Template With Provisions

BACKGROUND: We will be reformulating questions generated by non-lawyer users related to commercial law, business law, or similar areas.

TASK: We take a user question and reformulate it to ask the first, specific question that will help them ultimately answer the legal question they have. It is vital that the question accurately captures and/or maintains any legal issue conveyed by the user.

INSTRUCTIONS:

Given the user's question, formulate the first, specified question that should be asked based on the question provided by the user.
1. If you think the question does not need reformulating, you can simply return the question that was asked.

EXAMPLES:

Example 1:
User Question: Can I work somewhere else after I quit my job?
Reformulated Question: Could the terms and conditions of my present employment prevent me from working somewhere else if I quit?

Example 2:
User Question: What happens if the contract is thrown out?
Reformulated Question: What are the potential consequences of a contract or parts thereof being invalidated by court?

Example 3:
User Question: Who am I allowed to talk to about sensitive information?
Reformulated Question: Does the contract speak to confidentiality of information or non-disclosure?

EXPECTED OUTPUT:
The user's question was: [user question]
It has been reformulated to say: [reformulated question]

TASK INFORMATION:
The user's question is: {{USER QUESTION}}

GPT Prompt Template Without Provisions

Figure 4: Prompt templates given to GPT models with (left) and without (right) contract provisions to generate question reformulations.

```
<s>[INST]
BACKGROUND: We will be reformulating questions generated by non-lawyer users related
to commercial law, business law, or similar areas.

To help you understand the appropriate language you should consder we are going to
include provide some provisions from legal contracts that
have to do with commercial law, business law or similar areas.

These are complex legal provisions that cover a broad range or commercial/business
contracts.

Often, the user will ask a question that requires a number of other questions to be answered
first.  Rather than answering a question
that is unspecified, it would be helpful for the user to have their question reformulated if
necessary.

TASK: We take a user question and reformulate it to ask the first, specific question that will
help them ultimately answer the legal
question they have. It is vital that the question accurately captures and/or maintains any
legal issue conveyed by the user.

INSTRUCTIONS:

Given the user's question, formulate the first, specified question that should be asked based
on the question provided by the user.
1. To help you we have provided provisions from a contract based on their similarity to the
user's question.  We have provided you with the top results of the similarity assessment.
2. However, you are NOT to use the provisions to craft a response, you are to ONLY use
them to help reformulate, if necessary the first, specific question that a user should ask
based on their initial question.  There is in no way a guarantee that the provisions are on
point.  This is vital to understand.
3. If you think the question does not need reformulating, you can simply return the question
that was asked.

PROVISIONS:

{{CONTRACT PROVISIONS}}

EXAMPLES:

Example 1:
User Question: Can I work somewhere else after I quit my job?
Reformulated Question: Could the terms and conditions of my present employment prevent
me from working somewhere else if I quit?

Example 2:
User Question: What happens if the contract is thrown out?
Reformulated Question: What are the potential consequences of a contract or parts thereof
being invalidated by court?

Example 3:
User Question: Who am I allowed to talk to about sensitive information?
Reformulated Question: Does the contract speak to confidentiality of information or non-
disclosure?

EXPECTED OUTPUT:
To the provided user question respond with a single reformulation and nothing else. Return
the following output:

{
"user_question": "<Example_user_question>", \\
"reformulation": "<Example_user_question_reformulation>" \\
}

TASK:
The user's question is: {{USER QUESTION}} [/INST]
```

Llama/Mistral Prompt Template With Provisions

```
<s>[INST]
BACKGROUND: We will be reformulating questions generated by non-lawyer
users related to commercial law, business law, or similar areas.

TASK: We take a user question and reformulate it to ask the first, specific
question that will help them ultimately answer the legal
question they have. It is vital that the question accurately captures and/or
maintains any legal issue conveyed by the user.

INSTRUCTIONS:

Given the user's question, formulate the first, specified question that should be
asked based on the question provided by the user.
1. If you think the question does not need reformulating, you can simply return
the question that was asked.

EXAMPLES:

Example 1:
User Question: Can I work somewhere else after I quit my job?
Reformulated Question: Could the terms and conditions of my present
employment prevent me from working somewhere else if I quit?

Example 2:
User Question: What happens if the contract is thrown out?
Reformulated Question: What are the potential consequences of a contract or
parts thereof being invalidated by court?

Example 3:
User Question: Who am I allowed to talk to about sensitive information?
Reformulated Question: Does the contract speak to confidentiality of
information or non-disclosure?

EXPECTED OUTPUT:
To the provided user question respond with a single reformulation and nothing
else. Return the following output:

{
``user_question": ``<Example_user_question>",  \\
``reformulation": ``<Example_user_question_reformulation>" \\
}

TASK:
The user's question is: \{\{USER QUESTION\}\} [/INST]
```

Llama/Mistral Prompt Template Without Provisions

Figure 5: Prompt templates given to Llama and Mistral with (left) and without (right) contract provisions to generate question reformulations. Unlike the prompts given to GPT, these prompts contain explicit instruction tags.

be can be

E Annotator Agreement

To assess the robustness of our Legal Correctness scores, we calculate annotator agreement, specifically focusing on the legal correctness category, where a legal background is most crucial for making accurate judgments.

Agreement is measured across the entire set of expert evaluations. Cohen's κ suffers from a paradox where, if there is a high prevalence of one class the score can be too low even if annotators agree (Zec et al., 2017). Here, scores are predominantly 1 or 0, which could imply an issue with the κ score. Therefore, we include Gwet's AC1 in our analysis. This statistic also measures inter-rater agreement and accounts for chance agreement, but is not as sensitive to a highly dominant class (Wongpakaran et al., 2013). Based on the results in Table 6, the agreement among expert annotators is generally moderate. To calculate the scores of model performance, the scores of the evaluator most familiar with the legal domain were accepted.

To fully understand the importance of possessing a legal background when evaluating these prompts, we calculate the agreement between experts and non-experts for the legal correctness category. A non-expert author of the paper, unfamiliar with the domain problem, annotated a subset of the model's output. We observe very weak agreement between experts and non-experts, indicating that non-experts cannot effectively evaluate the output of a model with respect to legal correctness (Table 6).

	Between experts			Between experts and non-experts		
Model	Cohen's κ	% Agree	Gwet's AC1	Cohen's κ	% Agree	Gwet's AC1
GPT-4-Turbo	0.25	71%	0.52	-0.21	65%	0.51
Llama3-70B	0.32	67%	0.41	0.16	60%	0.20
Mistral-7B	0.16	66%	0.42	0.08	55%	0.12

Table 6: Agreement between expert annotators (left) and expert and non-experts (right) as to the legal correctness of the output. We include Gwet's AC1 in our agreement scores because Cohen's κ can sensitive to class imbalances. We find moderate agreement amongst experts for legal correctness, and observe much lower agreement between experts and non-experts