AMEND: Question Augmentation for Meaningful Exchange and Nuanced Discourse between experts and community members

Anonymous ACL submission

Abstract

It has become commonplace for everyday users to utilize large language models (LLMs) for day-to-day tasks. However, some of these tasks require specialized expertise, such as interpreting terms and conditions, calculating personal taxes, or isolating the likely cause of a medical symptom. Currently, users can ask these questions directly to LLMs and receive answers, but users may not possess the necessary expertise to instruct models and interpret their results.

In this work, we investigate whether models can output correct and factual answers when given expert-designed prompts within the legal domain. We propose the task of legal question reformulation, and develop a novel evaluation framework using expert legal knowledge. Through careful evaluation, we find that while LLMs show strong potential for legal reasoning, there are key safety issues present in their outputs. We also observe that adding linguistically-aligned in-domain text samples can improve performance, even when the samples are not aligned factually with the given question.

1 Introduction

Large language models (LLMs) have become ubiquitous in everyday life, with users prompting models to perform a wide variety of tasks. However, the text produced by these models is heavily dependent upon how they are prompted. For more general, everyday tasks, users may be able to prompt these models in a straightforward manner due to their own familiarity with the task. However, prompting often requires very explicit instructions and well-chosen examples. Tasks that require specialized knowledge, such as medical diagnosis, forming legal arguments, or interpreting contracts, may require years of training and formal education to perform effectively. Without an intricate understanding of how to perform these tasks, it is difficult to know how exactly to instruct the model or decide which details are most pertinent to the task.

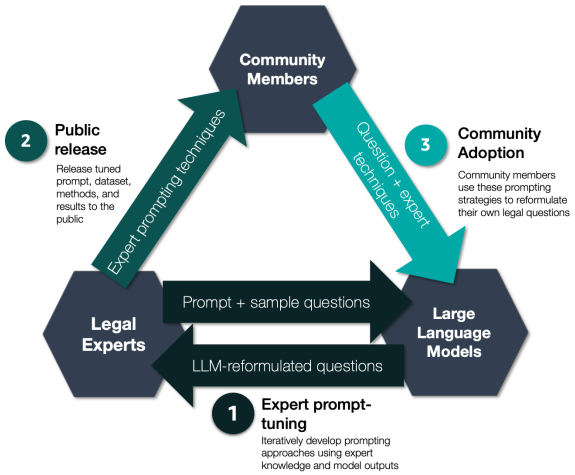

Figure 1: A visual representation of our proposed approach. We hypothesize that prompt templates developed by experts, along with relevant domain knowledge, can result in higher-quality outputs from LLMs for the task of legal question reformulation.

Thus, non-expert users attempting to gain expert insight from language models are at risk of receiving incorrect answers or unsafe advice in potentially high-stakes situations.

One way to close this gap, and to empower users to use these systems even without possessing the requisite expertise, is to enlist practitioners at the intersection of multiple disciplines. By recruiting researchers with experience with NLP systems and specialized training in other areas, we can design prompting strategies and templates that optimize LLMs' performance in those areas. Users can then use these strategies and templates for their own needs, and thus can make indirect use of expert knowledge and pertinent background information to get more accurate and useful answers to their questions.

The legal domain is an important use case for these strategies. People take legal actions every day. Some examples include signing: leases, employment agreements, terms of use, and store purchases.

However, they typically do not have a lawyer with them when they engage in these tasks; legal advice is expensive to obtain, as individually lawyers spend years in school learning their craft. Further, although lawyers are trained to "issue spot", or identify from factual descriptions what legal rights or obligations an individual may have, it is imperative that a non-lawyer seeking help be able to clearly communicate their question or circumstances. Misunderstandings between a lawyer and their client could lead to missed issues, inefficient performance, and frustration. Engagement between a lawyer and a community member often starts with a question. 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 community member ask an accurate question.

Herein, we address the novel task of reformulating a community member's question to better capture and communicate what legal issue they may have. This task is important given the prolific use of LLMs in every day life. For example, many online question forums employ an LLM such as ChatGPT to assist users. In our case study, we use questions posted to an online forum where users can ask lawyers a question. We choose to focus on questions concerning business/commercial law. We choose to address the task of question reformulation as these kinds of interactions often set the stage for interactions between experts and community members.

In this study, we bring together experts in law and NLP, and community members to provide a holistic approach to evaluating the use of LLMs for such a task. This paper is not about benchmarking different models and assessing their performance on a particular task. Rather, we introduce a method of prompting and analysis adaptable to domains where expert knowledge, and more effectively *integrating* expert knowledge, is crucial for meaningful performance. We illustrate the principles of this method, and its impacts on community members, in Figure 1. We then apply our method, utilizing the expert knowledge and skills of two lawyers who also conduct research in NLP and are authors on this paper. We create prompt templates for multiple different language models using extensive expert tuning. We supplement these tuned prompts by adding contextually-relevant contract clauses from the CUAD dataset (which contains commercial contracts) (Hendrycks et al., 2021) to the prompt, and describe our method for choosing these clauses in detail. We employ these strategies with several LLMs of varying sizes and access levels on publicly-available legal forum posts from real users. Our team members who are lawyers conduct a careful evaluation of their outputs, using a novel protocol developed by lawyers, introduced in this paper, we provide quantitative and qualitative analyses of each LLM's performance. Summarily, we address the following research questions:

RQ1 Can expert-tuned prompts correctly reformulate community member questions about complex, domain-specific issues?

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 How can experts evaluate safety concerns that would prevent an LLM's deployment on a particular task?

Following evaluation, although we achieve promising results, we ultimately conclude that these models are not suitable for safe deployment in a zero-shot setting. This is true even with state-of-the-art models and advanced prompting techniques involving expert tuning. We discuss our results and findings in detail and connect them back to the above research questions. 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 consisting of the generated reformulations from each of our baselines, along with code for retrieving the original data.

2 Related Work

Prompting Language Models The rise of LLMs has allowed users to *prompt* pretrained models with instructions for a given task, which often yields good results without requiring finetuning. LLMs are very large and difficult to interpret (and, in some cases, their weights and training details are not available to the public), but empirical prompting studies indicate that certain prompt formulation strategies may yield higher quality results. An example is chain-of-thought prompting (Wei et al., 2022), where the model is prompted to provide reasoning steps before giving its final answer. LLMs such as OpenAI's GPT models (Brown et al.,

2020) have been shown to perform well in zero- and few-shot settings, even for very specialized tasks (Agrawal et al., 2022; Savelka et al., 2023). However, in order to choose an effective prompting strategy for these specialized tasks, one must have intricate knowledge of how these tasks are formulated and carried out and be able to convey this knowledge in their instructions. We show that language models can achieve good results on the specialized task of legal question reformulation when prompted by domain experts, and release our prompting strategies to the public so that non-experts looking for legal advice can use our system with few barriers. We find that automatically picking linguistically aligned examples using a simple similarity score (§3) can improve results without the need for manual fact-checking or example selection. This further lowers the barrier for users without existing legal knowledge, who can select prompt examples from a text dataset within their specific legal domain using simple NLP methods.

Underspecified and Ill-Formed Questions Correctly answering a question requires understanding what exactly the question is asking. This becomes more difficult in the face of questions that are not well-formed (Faruqui and Das, 2018) or 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 deal with unclear questions. In this work, we focus on question reformulation for the legal domain, where asking the right questions often requires legal background.

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 task we propose is novel. In (Branting, 2017), although in the context of another problem, commented the difference in how laypersons and lawyers would pose a legal query. The authors notes that gaps in terminology and the complexities associated with legal concepts as a barrier between a lawyer's and layperson's understanding.

3 Data

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.

Legal Questions To compile legal questions posted in-the-wild by 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). Practicing lawyers registered to the site can reply with answers to these questions, and users can upvote or report these responses. In this work, we focus on contract law questions within the business domain. Thus, 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.

To gain some insight on the characteristics of our dataset, we plot a series of distributions based on the metadata provided by Justia in Figure 2 in Appendix A. We find that our questions most often pertain to California and Florida state law. A plurality of questions were posted in 2023, and questions had an average question length of 22.18 tokens.

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

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

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

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 user's 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. Some examples of user questions and corresponding linguistically aligned clauses can be found in Table 2.

4 Experiments

In this work, our goal is to determine whether LLMs can produce high-quality question reformulations when assisted by expert prompting and relevant domain knowledge. In the following, we describe our prompting techniques, baselines, and experimental settings.

Expert Prompting To bridge the knowledge gap between community members and experts, we can consult domain experts with a background in NLP and language modeling, who can create prompt templates that instruct language models to perform a given task while providing them with relevant background information. Making these templates and methods available to the public allows everyday users to utilize this expertise without years of study or practice. In our case study, we address the task of question reformulation in the legal domain, so that users looking for particular answers or information can form specific and relevant questions using appropriate language.

To formulate our prompts, we utilize the unique expertise of a lawyer studying NLP (an author on this paper). Generally, the prompts were developed using iterative approach to prompt formulation by manually inspecting outputs for correctness and adjusting the prompt accordingly. However, the core components necessary for an expert tuned prompt to achieve suitable results are:

- Briefly clearly describe the area, domain, or problem relevant to the scenario. In this case it is the proper formulation of issues in the legal domain, specifically, business/commercial law.
- Specify the narrow domain specific task that the model will be performing. Here, that is reformulating a community member's question to remedy implicit and under-specified characteristics, while maintaining the user's original intent.
- Explain any supplemental knowledge included in the prompt and its potential utility and risks, providing specific rules for the use of the supplemental knowledge.
- Provide ideal in-context examples of how the model should perform.

The idea 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 our prompts are provided in full in §B.

Providing Relevant Legal Language In addition to expert-guided prompting, we experiment with incorporating contract provisions deemed similar to the 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 B. Examples of the most similar clauses to selected user questions are shown in Table 2. 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.

Models and Experimental Settings To test our hypotheses, 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 B. We first run OpenAI's GPT-4-Turbo (`gpt-4-0125-preview`) (Achiam et al., 2023) at a temperature of 0.0, with

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 2: Sample questions from our dataset, along with the clause deemed most relevant to the question based on the cosine similarity score.

a maximum of 250 new tokens, a top p value of 1, a frequency penalty of 0, and a presence penalty of 0. As explained in (Gray et al., 2023; Savelka et al., 2023), these parameters are used to obtain deterministic output, which is vital in the legal domain.

We then run two open-source language models: Meta's LLAMA 2 70B chat model (`Llama-2-70b-chat-hf`) (Touvron et al., 2023a,b), with 70 billion parameters, and Mistral (Mistral-7B-Instruct-v0.2) (Jiang et al., 2023), with 7 billion parameters. We 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 C.

We choose these three models because they differ greatly in size and accessibility. GPT models can be accessed via a GUI online or through a mobile application. Thus, these models are accessible and easy-to-use for users for whom writing code and locally hosting models is a barrier, and certain GPT models are free to interact with via GUI. For these reasons, we include a GPT model in our baselines. However, making API calls or prompting specific GPT models through the online interface costs money, limiting access to those who are willing and able to pay to use GPT. Further, GPT is a closed model with stochastic outputs, and its weights and training details are not available to the public. It is also unknown how often GPT's weights are updated from additional training. Thus, GPT's results are not reproducible, and as such its research utility is limited. Because of this, we run two open-source LLMs of varying sizes in addition to GPT. Because researchers can gain access to their weights, which are not propriety, these models are much more valuable for research purposes. Further, they are free to run locally. We report results for Llama 2 70B, which has 70 billion parameters, and Mistral, which has 7 billion parameters, in order to provide useful benchmarks for users or practitioners with varying levels of compute. An important design constraint is to keep our prompting to zero-shot or in-context learning. This is to dispense with the time and expense needed to generate even a reasonably sized corpus of training data and to remain flexible as to the specific task needed by a practitioner. It is much easier and less costly to modify a prompt rather than reannotate data. We choose a number of different models to show performance on the task so that users with different resources have an idea of what to expect. We do not do this simply to benchmark model performance, but rather as an exploratory study regarding the characteristics of the models' reformulations.

5 Results

Here we discuss the results and their implications for the research questions identified in the introduction. We also introduce our expert-based evaluation criteria.

5.1 Evaluation Criteria

A key component of our pipeline is the steps used to evaluate the models from the perspective of the domain, which in this case is law. We also analyze the potential harm or safety issues presented in this task, both generally and domain specific. The eval-

uation of the model is based on the model's output given our two different styles of prompts. The first is the expert prompt without any relevant provisions included. The second is the expert prompt with linguistically aligned contractual provisions based on the user's question. To evaluate the model from a legal perspective, we use three criteria: legal correctness, factuality, and clarity/understanding. We define them as follows:

> **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?
>
> **Factuality**: Which reformulation most closely aligns factually with the original question?
>
> **Clarity/Understanding**: 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 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 both reformulations are equally satisfactory, both are scored as 1. If both reformulations are equally unsatisfactory, both are scored as 0. The quantitative evaluation was conducted by two lawyers, and we report moderate agreement using Cohen's Kappa (McHugh, 2012) and Gwet's AC1, which is less sensitive to dominant classes (Wongpakaran et al., 2013). Additional details, including the agreement scores for each model, can be found in Appendix D.

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

5.2 Performance

To evaluate each model's performance across the different categories, we randomly sampled a subset of questions (the same subset was used for each model) for evaluation. During evaluation, the prompt (with or without relevant provisions) from which the question was generated was unknown to the evaluators. We examine the performance of each model on each category in Table 3.

Metric	With Provs	Without Provs
GPT-4-Turbo		
Legal Correctness	0.65	**0.75**
Factuality	0.84	**1.0**
Clarity	0.94	**1.0**
Llama2-70B		
Correctness	**0.52**	0.41
Factuality	**0.73**	0.58
Clarity	**0.63**	0.57
Mistral-7B		
Correctness	**0.37**	0.31
Factuality	0.53	**0.59**
Clarity	**0.53**	0.39

Table 3: Because individual scores are either 0 or 1, when averaged, scores closer to 1 are better than scores closer to 0.

6 Discussion

We ultimately conclude that although our results are promising, even the model with the most parameters (GPT-4-Turbo) is not capable of solving the task with suitable quality for deployment. Moreover, the models' output raise a number of concerns with respect to safety. With respect to RQ1, we conclude that expert-tuned prompts show a reasonable capability in reformulating a user's domain specific question. As for RQ2, with the exception of GPT-4-Turbo the inclusion of linguistically relevant information influences a model's performance in a strongly positive manner. As for RQ3, all models display safety concerns, that unless remedied are a barrier to deployment.

6.1 Quantitative Evaluation

It is expected that correctness would have the lowest scores compared to the other categories. This is because evaluating legal correctness requires assessing the legal issues raised in the reformulation, while the other categories are concerned more with language quality. As legal reasoning is much more difficult for language models than producing quality, factual language, this behavior makes sense. Being able to identify a potential legal issue is a skill necessary for lawyers (this is one of a lawyer's most important skills), and it can be difficult, especially for lawyers that do not have experience in a particular area. We believe our results provide

evidence that the models can, given our prompting framework, exhibit some level of legal reasoning.

First, to assess the robustness of our results as to legal correctness, in Appendix D we present detailed results about annotator agreement. Generally we find moderate agreements amongst experts, indicating our results are reliable. As a way to assess risks of deployment, we incorporate community member knowledge by calculating agreement between experts and non-experts as to legal correctness and find weak annotator agreement, indicating that non-experts are not able to evaluate legal correctness effectively. With respect to all models investigated, GPT-4-Turbo had the best overall performance both with and without provisions included in the prompt. For this model, it seems that the inclusion of relevant prompts degrades the model's performance across all categories. The biggest disparity comes with the factuality of the model. Here, we see a .16 decrease in performance. This is particularly concerning, as factual discrepancies can result in inaccurate legal analysis, or issue identification. As for `Llama2-70B`, the inclusion of provisions in the prompt seems to result in a strong increase in performance across all categories. The same is true for `Mistral-7B`, except with the factuality category.

Evaluators report that the differences in the reformulated questions using GPT-4-Turbo with or without provisions is generally very slight. Almost always the difference comes down to the use of a single word that makes one reformulation better or worse than the other. As for `Llama2-70B`, the differences are more pronounced and it was easier to identify a better reformulation. Often, `Mistral-7B` output unusable reformulations that were nonsensical or were no different from the user's original question.

6.2 Qualitative Analysis

A review of the qualitative notes provided by evaluators reveals important aspects of the a model's output. For GPT-4-Turbo, often, the model outputs an identical or near identical reformulation for both prompting schemes. This indicates that the model may be insensitive to user instruction, or perhaps a response is being retrieved rather than generated probabilistically. This may also explain why including linguistically aligned contract provisions caused the model's performance to decrease. We provide examples of model reformulations in Table 4, and describe a few examples in detail below.

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:

> *Can I get sued at 15 for money?*

Every model answered the question. Ideally, none of the models 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 strong legal reasoning. Although the models should not have answered the above question, GPT-4-Turbo's formulation displayed promising legal reasoning:

> *Are minors legally liable for financial obligations or debts, and under what circumstances can they be sued for money?*

Although subtle, there is a major difference between being liable for a financial obligation and debt, and being sued for money generally. The model picked up on this difference and appropriately used the terms.

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.

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 also counsel their clients beyond legal action, maybe legal action isn't the best thing. So when they say financial strategies, the model is correctly implying a non-legal option could be viable. See Model Rules of Professional Conduct Rule 2.1"*

Table 4: Example questions and their reformulated counterparts output by LLMs. Comments from the evaluators regarding the effectiveness or safety of the reformulated question are included.

Models sometimes reformulate questions requiring immediate legal attention. All models also answered sensitive questions that warrant a particular, strong response from an attorney. For example:

> *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 limitation: 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.

Models sometimes inject unnecessary jargon into their reformulations. In some cases, `Llama2-70B`, would output unnecessary legal jargon:

> **User Question**: *Can a consultant sign documents on behalf of their client?*
> **Reformulation**: *Does the contract or agreement provide provisions for the consultant to act as the client's agent or attorney-in-fact for the purpose of execute documents on their behalf?*

First, the question asked by the user is likely sufficient on its own. The model ignored the prompt instructions stating that acceptable questions could go un-reformulated. Second, when the model reformulated the question, it added unnecessarily the terms "client's agent" and "attorney-in-fact". 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 both `Llama2-70B` and `Mistral-7B` 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. Calling back to the scenario in our introduction, if any of these models were available on a website to perform such a task, serious safety concerns would arise.

7 Conclusion

Given the prolific rise and use of LLMs in an every day setting, often without expert supervision, even seemingly simple tasks could result in harmful outcomes. Including expert knowledge could potentially result in better performance and safety. In this work, we introduce a novel question reformulation task and explore three key research questions. We first show that zero-shot expert-tuned prompts show reasonable performance in reformulating community member questions to more clearly communicate potential issues (**RQ1**). But even the strongest model is legally incorrect in $\frac{1}{4}$ of our instances. Further, we find that the inclusion of linguistically relevant language in the prompts strongly improves the performance of open source models (**RQ2**). However, the same technique cause GPT-4-Turbo to decrease in performance. Given the nature of the model, the cause of this is difficult to investigate. Lastly, we provide a paradigm for how experts can criticize and evaluate a model's output, particularly for safety concerns (**RQ3**). From this analysis, we learn that there are significant barriers to safety before these models can be deployed in this setting.

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 create an attorney client relationship. The questions relied upon to annotate data are public available. We do not work with sensitive or personal data. We have not explicitly shared any data with OpenAI. 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. Users of this framework should think carefully and consult with a legal expert before deploying this framework on a discrete task. The data that was used in this research was obtained and used within the scope of permission of the authors in **?**.

References

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

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*, volume 34, 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*.

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, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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 2 provides some insight into our dataset of 247 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 Prompt Templates

B.1 GPT

With contract provisions The following template 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.

> 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

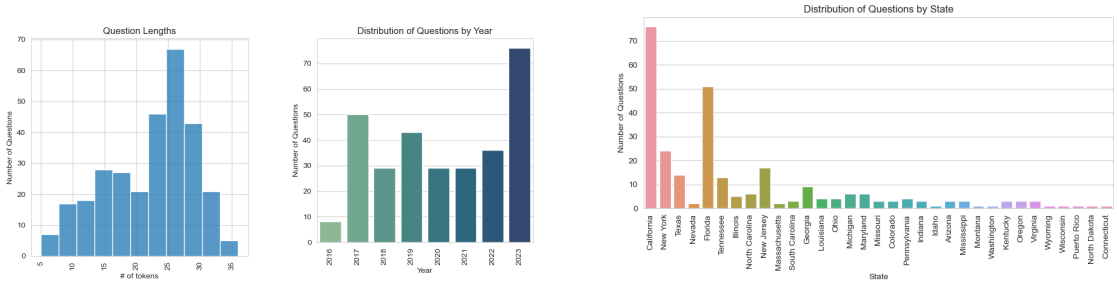

Figure 2: Distributional plots for our dataset of 274 Justia questions. The left plot depicts the distribution of question lengths, the center plot counts the number of questions asked each year, and the rightmost plot depicts the number of questions from each state in our dataset.

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}}

Without contract provisions The following template was provided to the language model to study its performance when relevant contract provisions are not added to the prompt. `{{USER QUESTION}}` is replaced with the title of the question posted on Justia.

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 infor-

mation? 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

MC: Fill in these sections

B.2 Llama/Mistral Prompts

With provisions prompt = """<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] """

Without 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]"""

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

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 be can be

D Annotator Agreement

Lastly, we assessed annotator agreement amongst experts as to the scores given for the legal correctness category. Agreement was 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, in many of the evaluation scores were 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 5, the agreement amongst 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 assess whether non-experts are capable of evaluating the legal correctness of a model's output, a non-expert author of the paper, unfamiliar with the domain problem annotated a subset of the model's output. As shown in Table 6 the agreement between experts and non-experts is very weak, indicating that non-experts cannot effectively evaluate the output of a model with respect to legal correctness.

Model	Cohen's κ	% Agreement	Gwet's AC1
GPT-4-Turbo	0.25	0.71%	0.52
Llama2-70B	0.32	0.67%	0.41
Mistral-7B	0.16	0.66%	0.42

Table 5: Agreement between expert annotators as to the legal correctness of the output.

Model	Cohen's κ	% Agreement	Gwet's AC1
	Expert-Non Expert Agreement		
GPT-4-Turbo	-0.21	0.65%	0.51
Llama2-70B	0.16	0.60%	0.20
Mistral-7B	0.08	0.55%	0.12

Table 6: Agreement between expert-non expert annotations as to the legal correctness of the output.