Experimenting with Legal AI Solutions: The Case of Question-Answering for Access to Justice

Jonathan Li [1] **Rohan V Bhambhoria** [1 2] **Samuel Dahan** [1 3 4] **Xiaodan Zhu** [1 2]

Abstract

Generative AI models, such as the GPT and Llama series, have significant potential to assist laypeople in answering legal questions. However, little prior work focuses on the data sourcing, inference, and evaluation of these models in the context of laypersons. To this end, we propose a *human-centric* legal NLP pipeline, covering data sourcing, inference, and evaluation. We introduce and release a dataset, LegalQA, with real and specific legal questions spanning from employment law to criminal law, corresponding answers written by legal experts, and citations for each answer. We develop an automatic evaluation protocol for this dataset, then show that retrieval-augmented generation from only 850 citations in the train set can match or outperform internet-wide retrieval, despite containing 9 orders of magnitude less data. Finally, we propose future directions for open-sourced efforts, which fall behind closed-sourced models.

1. Introduction

Today, much of natural language processing rests on *high-quality data*. Advances in unstructured data for pretraining have allowed general language models to be more lightweight, performant, and therefore accessible (Abdin et al., 2024; Team et al., 2024). Language models are promising in providing legal advice to laypeople, but prior work suggests that they struggle with hallucination (Dahan et al., 2023). Inspired by the high-quality data that powers general domains, we identify a gap in high-quality structured legal data (e.g., question-answer pairs) approved by legal experts. In this work, we hope to build more *human-centric legal AI* systems by improving the data source to address laypeople. We use this data at *retrieval time* to improve model performance, which does not require additional training or fine-tuning.

Little prior work focuses on optimizing legal AI systems from start to finish for factors that matter to laypeople. Among these factors are accessibility of the services due to cost, factual correctness, and ease of understanding. In this paper, we propose an end-to-end *human-centeric legal AI* framework, which covers data sourcing, training/inference, and evaluation to improve these factors; importantly, we put laypeople first by ensuring each step of the process is backed by high-quality data from legal experts (see Figure 1). To our knowledge, this type of *human-centric legal framework* is the first of its kind.

First, we construct a high-quality evaluation dataset of 323 questions asked by laypeople on real legal questions and answers vetted by legal experts. We ask law students to write expert answers to these questions and release this dataset to the public. Then, we develop an automatic evaluation protocol based on the *factuality* of the generated answer, as a legal expert would. Inspired by massive improvements to model quality through higher quality data at *training time* (e.g., Phi-3; Abdin et al., 2024), we improve the data sourcing process at *retrieval time*. Specifically, we propose domain-specific retrieval, bolstering the performance of existing LLMs on legal question-answering by retrieving from sources trusted by legal experts. We show that retrieval from under a thousand legal-expert-approved articles matches or exceeds the performance of retrieval from hundreds of millions of internet articles.

Overall, our contributions are as follows:

- We construct a dataset containing real legal questions and high-quality answers labelled by legal experts. We release the evaluation dataset publically.
- We create an evaluation protocol vetted by legal experts and find that existing models have room for improvement in factuality.
- We show that domain-specific retrieval from relatively

[1] Ingenuity Labs Research Institute, Kingston, Canada [2] Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada [3] Department of Law, Queen's University, Kingston, Canada [4] Cornell Law School, Cornell University, Ithaca, United States. Correspondence to: Xiaodan Zhu <xiaodan.zhu@queensu.ca>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

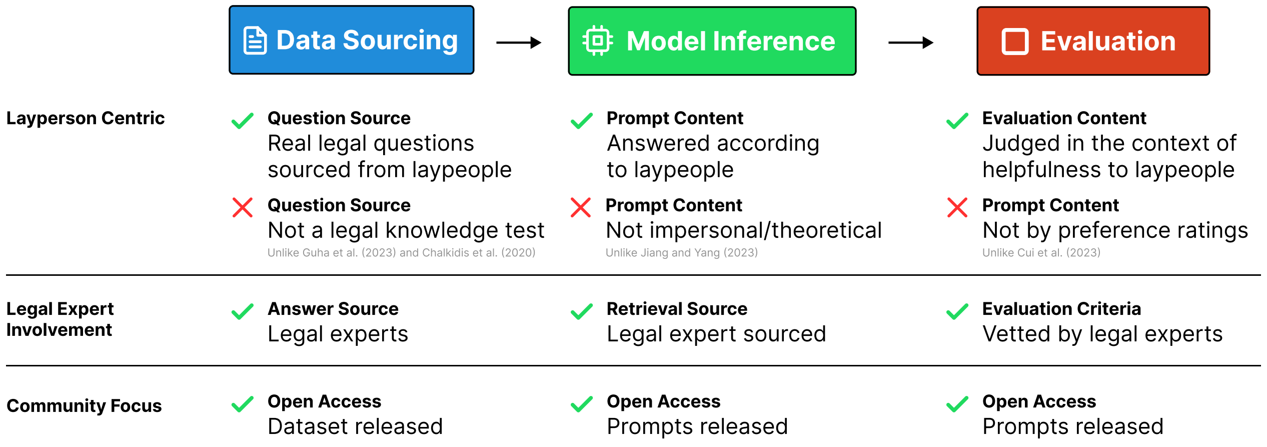


Figure 1. An overview of our framework for human-centric legal AI.

few sources trusted by legal experts can outperform existing non-retrieval models and match or exceed retrieval-augmented models that rely on the entirety of the internet.

2. Prior Work

Retrieval Augmented Generation. Large language models often hallucinate and contain outdated information (Zhang et al., 2023). Retrieval augmented generation (RAG) is an emerging approach to reduce the prevalence of hallucinations by grounding a model's generations in a data source besides the model's weights. Retrieval has been used extensively in single-hop (Ke et al., 2024), multi-hop (Sun et al., 2023), and long-form open-ended question answering (Lin et al., 2023). With the rise of instruction-following language models (Touvron et al., 2023; Chung et al., 2022; Brown et al., 2020), retrieval methods often insert context directly into the context of the language model (Ma et al., 2023; Chen et al., 2023). We focus on this setting because it is possible to integrate with existing well-performing models (such as OpenAI's GPT-3.5), further supporting our human-centric goal of making legal AI more accessible.

Very recently, commercial RAG efforts such as Cohere's `Command R+` models, have been applied to legal domains with a focus on trustworthiness and data privacy (Gainer & Starostin, 2024). Their retrieval method passes the retrieved documents directly into the context of a language model, such that the model generations are grounded in the context provided. In this work, we build off this line of research by focusing on retrieval from a trusted source.

Datasets and Legal AI Benchmarks. NLP has been applied to various fields in law, such as question answering, relation extraction, or text summarization (Zhong et al., 2020). Previously, work was focused on domain-specific fine-tuned models (Chalkidis et al., 2020; Zheng et al., 2021). Recently, existing work has focused more on the ability of *general LLMs* to perform legal reasoning (Yu et al., 2022; Jiang & Yang, 2023; Blair-Stanek et al., 2023; Yu et al., 2023). Regarding benchmarks for LLMs in legal applications, existing benchmarks exist in Chinese (Duan et al., 2019; Dai et al., 2024) and American (Guha et al., 2023) law. However, many existing benchmarks lack evaluation of open-ended responses which are of interest to laypeople who ask language models for legal advice directly.

We source our legal questions from a public legal advice forum (though answers are written in-house by legal experts). These online forums have been used extensively as sources of data for machine learning. For instance, Yao et al. (2020) uses data from a mental-health advice sub-community on Reddit (known as a "subreddit") to detect suicidality in opioid users. Li et al. (2022) uses the "r/legaladvice" subreddit for classification of Reddit posts in evaluation. We extend this work on forum-based data by considering a much more difficult task: generating a factually accurate legal answer to a given legal question.

Better Data for Better AI. Existing work has found that sourcing better data can lead to model improvements in the pretraining stage. Eldan & Li (2023) create a high-quality machine-generated dataset, then shows that very small models (order of tens of millions of parameters) can learn perfect grammar from this high-quality data. Taking this a step further, Li et al. (2023) and Gunasekar et al. (2023) introduce the Phi series models, which gain impressive performance despite their small size, driven by high-quality data. Very recently, open-sourced state-of-the-art language models Llama 3 (AI@Meta, 2024) and Phi-3 (Abdin et al., 2024) build off this idea, reaching state-of-the-art performance by improving data sourcing. Inspired by this work, we review a related but orthogonal direction: improving the reliability of data sourcing at *retrieval time*, rather than just at *pre-training time* like in prior work.

Automatic Evaluation. As LLMs become better at problem-solving, their potential to evaluate responses relative to a gold answer becomes increasingly attractive (Oh et al., 2024). Current work has focused on evaluation of open-ended model generations for general domains, such as trivia question answering (Wang et al., 2023) or everyday conversational settings (Lin & Chen, 2023). In legal AI, however, most models are evaluated on closed-ended tasks that can be trivially graded (Koniaris et al., 2023; Xu & Ashley, 2023; Savelka, 2023). Cui et al. (2023) evaluates model generations using crowd-sourced human preference ratings in Chinese (with an ELO system), a step in evaluation for open-ended generations. Bhambhoria et al. (2024) explores the possibility of automatic evaluation in the legal domain, showing that most classified samples align with the expert opinion. In this work, we aim to capitalize on the benefits of automatic evaluation while optimizing the process for legal factuality evaluation by consulting with legal experts.

3. Methods

To build a source of structured and expert-approved legal data that is effective at *retrieval time*, we construct a new dataset from real legal questions and ask law professors and law students to answer these questions. Then, we use this data during our retrieval process to ground model answers in citations vetted by legal experts. During the evaluation process, we also ground our evaluations with this dataset, establishing an end-to-end legal-expert-driven framework (see Figure 1).

Specifically, we source questions from an online community[1], collected from January 2021 to October 2022. These questions are specific (e.g., Table 2) situations that real laypeople have, not hypotheticals[2]. For instance, the example sample in Table 2 outlines a specific scenario. This real-world focus allows our evaluations to be closer to a target domain that is helpful to laypeople. Then, we ask law professors and law students to provide golden answers to these questions. Since this research was done in Canada, the legal experts we worked with were knowledgeable primarily in Canadian law. Therefore, we asked these annotators to answer these legal questions according to Canadian law. Human answers are typically concise (shorter than the questions) and under 100 tokens (see Figure 2). Each answer contains a citation with more information relevant to the question. To perform a rigorous performance analysis across legal areas of practice, we classify each question into six categories relevant to laypeople, shown in Table 1. The classification was done through a zero-shot classification approach (Laurer et al., 2023) and manually inspected for correctness. To aid the community in evaluating existing LLMs, we release our evaluation dataset ($n = 323$) publicly[3].

Table 1. Composition of the area of law for each question in our dataset.

Category	Percent
Employment and labour law	27.9
Family and juvenile law	27.1
Real estate law	21.4
Corporate law	9.2
Personal injury law	9.2
Civil rights law	5.2

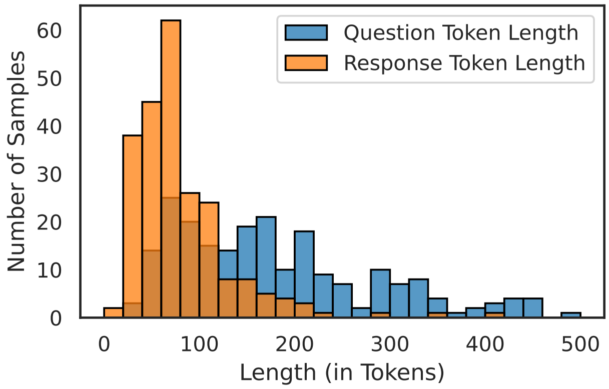


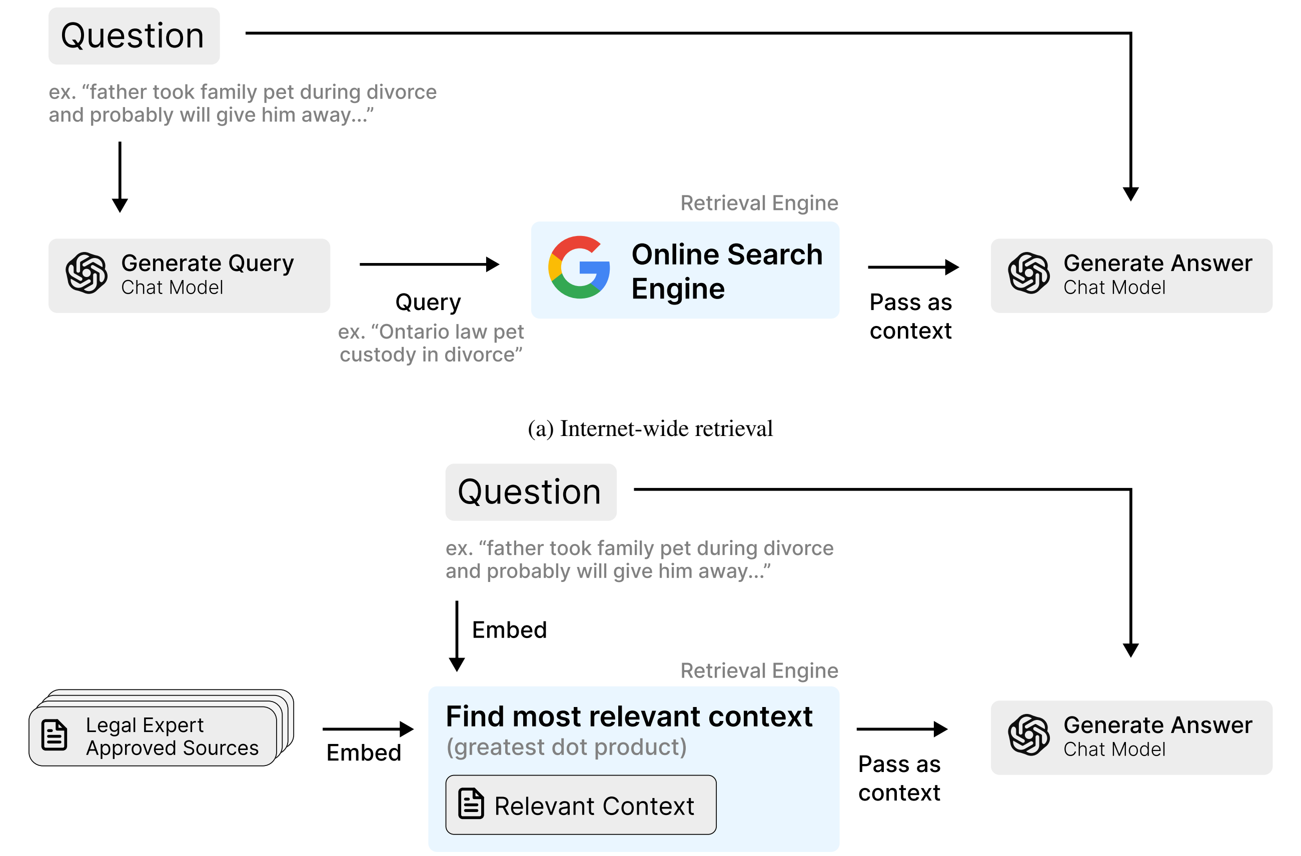
Figure 2. Distribution of question lengths and response lengths. Responses are concise and specific.

Inspired by Lin & Chen (2023), we employ human-grounded automatic evaluation of the generated answers. We use golden labels written by legal experts as ”grounding” for the language model, then ask the LLM (`GPT-4-0613`) to rate the answer’s factuality relative to the expert answer. If there is a factual contradiction with the expert answer, we treat the model response as factually disagreeing. We run the automatic evaluator with a temperature of zero (since we only need the prediction ”factual” or ”not factual”). In our study, we minimize this factual disagreement to build factually accurate models. This approach is among the criteria that legal experts use to evaluate answers by law students (Bhambhoria et al., 2024).

[1] https://www.reddit.com/r/legaladvice/

[2] As per the legal advice community guidelines, the questions must be real questions, not hypothetical questions

[3] https://huggingface.co/datasets/jonathanli/law_qa_eval


Question
ex. "father took family pet during divorce and probably will give him away..."
Generate Query
Chat Model
Query
ex. "Ontario law pet custody in divorce"
Retrieval Engine
Online Search Engine
Pass as context
Generate Answer
Chat Model
Question
ex. "father took family pet during divorce and probably will give him away..."
Embed
Legal Expert Approved Sources
Embed
Retrieval Engine
Find most relevant context
(greatest dot product)
Relevant Context
Pass as context
Generate Answer
Chat Model

(a) Internet-wide retrieval

(b) *Human-centric* legal retrieval **(ours)**

Figure 3. Retrieval-based methods used for our experiments. Given a legal question, retrieval is performed to generate a relevant answer.

Retrieval Methods. In addition to using existing language models to answer legal questions, we also consider grounding a model's responses in the context of a relevant article. As previously discussed, we believe the retrieval process could also benefit from high-quality legal data. In this work, we try to retrieve from only trusted legal sources, as shown in Figure 3b. Since we split the dataset into a train and test set, we use the citations from the train set ($n = 850$) as a set of trusted legal documents and only retrieve from these documents. This offers two main benefits: (a) the documents used by the model are known to be factual and helpful by legal experts, which is not the case for any document on the internet, and (b) searching a smaller subset of legal-expert-approved documents provides computational and storage benefits.

As shown in Figure 3b, we embed both the context and the question using an existing state-of-the-art embedding model, `BAAI/bge-large-en-v1.5` (Xiao et al., 2023). Then, we compute the dot product between the question and each document, selecting the document with the greatest dot product as the most relevant sample. Then we provide this document in the context of an existing language model (`GPT-3.5-turbo`), using prompts containing both the context and question. Unlike prior work, we evaluate retrieval from only legal expert sources rather than the entirety of the internet. This constitutes our model inference part in Figure 1. We call this "legal retrieval".

As a baseline, we evaluate (a) GPT-3.5-turbo without retrieval augmented generation and (b) GPT-3.5-turbo with retrieval from the entire internet. For (b), we use a language model to produce a query for a legal question that can be queried for in a search engine (Figure 3a). Then we use an existing web search engine (Google) to find the most relevant article and inject it into the context of the language model while answering this question. This article is used to provide context to the model. We call (b) "internet retrieval".

Internet retrieval is not as simple as retrieval from only legal documents, since a Google search is performed. Search engines likely contain more sophisticated methods than a simple embedding similarity check. When evaluating the performance of retrieval from the entirety of the internet

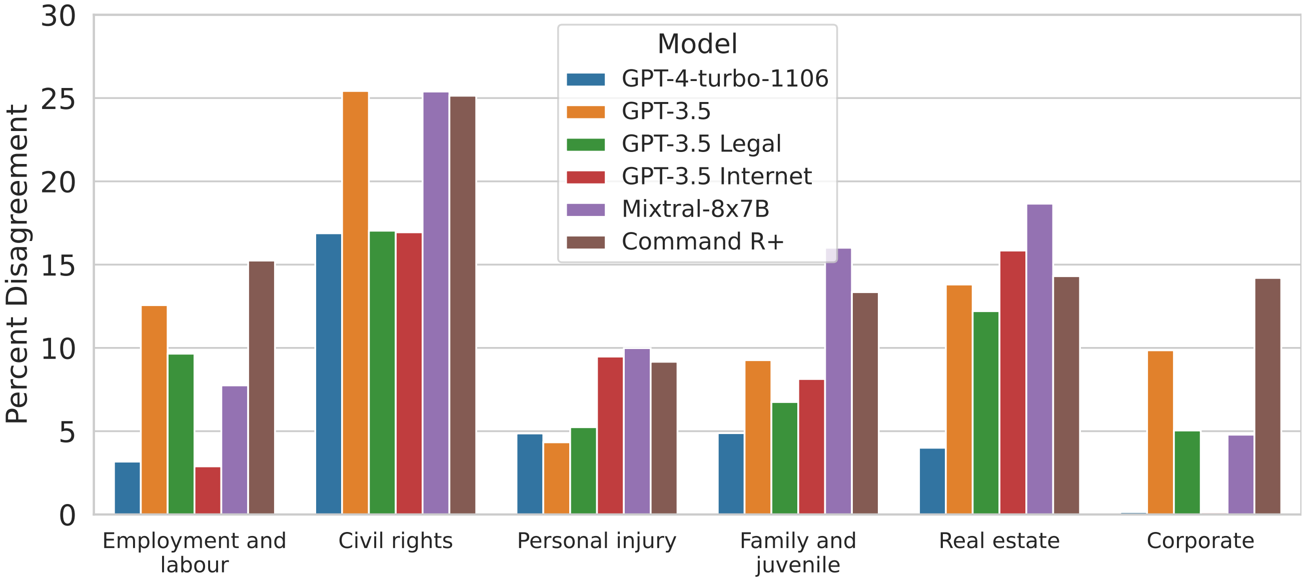


Figure 4. Factual disagreement of each model by category. Lower is better.

Table 2. Example question ("source") and provided answers and citation.

Source	Father took family pet during divorce and probably will give him away. My parents recently got a divorce a few months back and my father has been sending very disgusting harassing messages any way he can. He already has escapee warrants and now he is also getting more added to him due to his messages. He took the family dog who is legally under my mothers name and we know he is not safe with him. We have no idea where he is at but we know he is not a very safe person with our dog. We really fear of him giving the dog away to his friend and we wouldn't see him ever again. What do we do if that's the case? Can we fight to get him back if he gives him to his friend?
Answer	In Ontario, dogs are considered personal property. In determining which spouse has a right to the dog, a court will consider ownership papers as well as several factors such as: Is the pet more bonded to one person over the other? Who can best provide continued care? Who paid for the pet?
Citation	`https://www.siskinds.com/pet-custody-laws-in-ontario`

against retrieval from only legal documents, more computation occurs using an internet-wide search.

Generative Baselines. As a further baseline, we evaluate the state-of-the-art open-source models on this legal task. We use the state-of-the-art `Mixtral-8x7B` (Jiang et al., 2024) for non-retrieval generation and the state-of-the-art Cohere Command R+ model (Aiden, 2024) for retrieval-augmented generation. We rely on Cohere's pipeline for retrieval-augmented generation, whose model is purposefully tuned for this purpose.

For each method, we pass three samples from a separate train split as fewshot examples into the input prompt. We use the default generation and sampling settings (i.e., temperature and top-p) for each model.

4. Results and Discussion

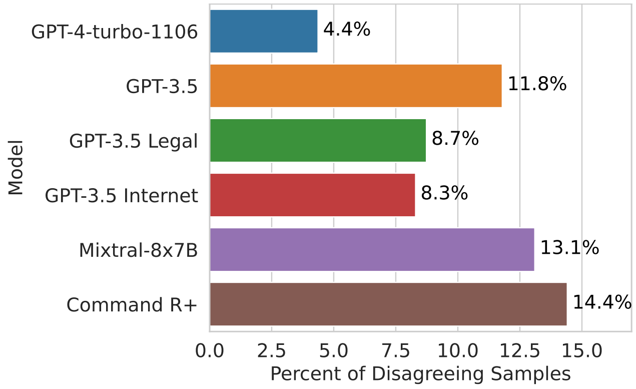


Figure 5. Factual disagreement for each model. "GPT-3.5 Legal" is retrieval using only legal documents, and "GPT-3.5 Internet" is retrieval from the entire internet.

As seen in Figure 5, the retrieval-based approaches tested typically perform better than their tested non-retrieval counterparts. Additionally, we make various observations:

Open-source models fall behind. Mixtral-8x7B and Command R+, state-of-the-art open language models perform worse than current closed-source models. Future work should continue testing the newest state-of-the-art open-sourced models for this task. Though we recognize the importance of open-source models, we also note that practically, the inference costs associated with using open-sourced models are nonzero for a layperson. Therefore, closed-sourced models like GPT-3.5 are still appealing from a cost perspective compared to other open-sourced models.

Using limited legal documents is just as useful as the entire internet. Retrieval from limited legal sources (just 850 documents) performs similarly to searching the entirety of the internet (hundreds of billions of documents). In the case of Cohere Command R+, a state-of-the-art enterprise retrieval system used in law [4], our simple retrieval method improves performance, despite Command R+ performing retrieval across the entire internet. The indexable internet contains more information than just our limited legal information, but the vastness of the space also make retrieval of documents much more difficult. Surprisingly, narrowing down the number of retrievable documents by 9 orders of magnitude[5] retains enough information to perform similarly to retrieval from the entire internet. By reducing the number of documents that can possibly be retrieved, human-driven review of individual documents is more feasible, and storage and computational costs decrease.

GPT-4 outperforms retrieval. We find that `gpt-4-turbo-1106` outperforms even retrieval models. However, we note that use of GPT-4 often has prohibitive costs (in some cases, 60x more), especially when paired with retrieval-augmented generations which span multiple contexts. Therefore, we still believe that relative improvements to GPT-3.5-turbo, a much more cost-effective model, continue to benefit the layperson despite the existence of more expensive and well-performing models.

Some categories of questions are more difficult to answer accurately. We find that some categories, such as "civil rights" and "real estate" are the most challenging for existing language models, illustrated in Figure 4. Qualitatively, we observed that questions falling under these areas of law contained the most specific and personal questions, and also had more nuanced cases (e.g., a highly specific question about a tenant's landlord).

In what situations does legal retrieval help? Apparent in Figure 6, in some cases retrieval from the entire internet can provide a lack of nuanced information, making the response factually incorrect compared to retrieval from only vetted legal documents. In the specific example presented, the retrieved web article failed to capture nuance when describing the punishment for damaging a vehicle during a traffic infraction, while the retrieved legal article from a vetted legal source did.

Figure 4 shows the strengths and weaknesses of each approach by category. Using retrieval from our expert source of legal documents is almost always better than or similar to the performance of non-retrieval methods. Additionally, retrieval from vetted legal sources performs on par with or better than retrieval from the entire internet, except for "employment and labour" and "corporate" categories. Further experiments are required to investigate the cause of this disparity, though we hypothesize it is because these categories have questions that span a larger space of questions, implying that retrieval from the internet is more applicable.

4.1. Future Directions

Closing the gap between open-sourced and closed-sourced models. As shown in Section 4, the gap between the top open-sourced model and the closed-sourced models (`GPT-4`) is substantial. From the perspective of human-centric legal AI, this is problematic as these black-box models often lack accountability in their data sources, which is especially important in the legal domain (Dahan et al., 2023).

Continual Updating. Laws and regulations constantly evolve, and legal AI systems need to stay up-to-date with the latest changes. Investigating techniques that can efficiently integrate new legal information and adapt models accordingly would better reflect the dynamic nature of the legal landscape. Currently, our methodology focuses on retrieving from a static set of documents, though the set of documents could be continually pruned and updated (Bhambhoria et al., 2024).

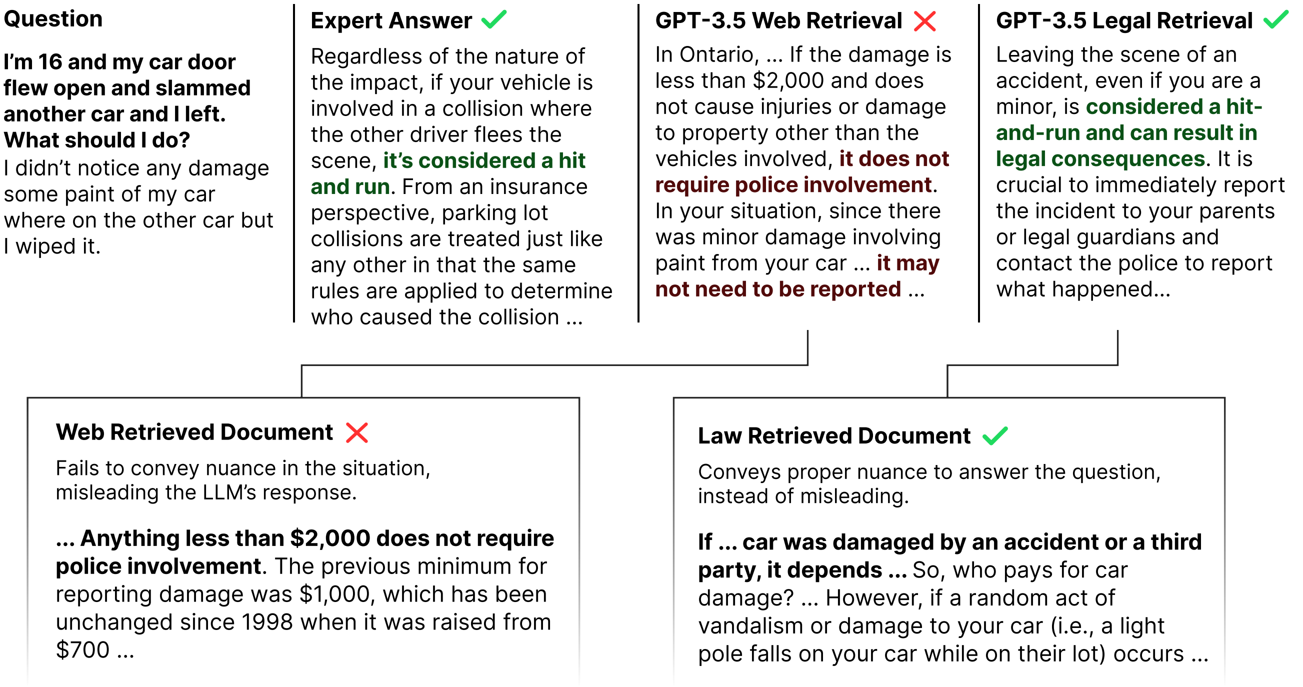
Unstructured Legal Data. In this work, we focus on sourcing high-quality structured data for use during retrieval and evaluation. Expert involvement in the data selection process could extend to unstructured data during the pre-training phase for the legal domain, which has already shown promise in general domains (Gunasekar et al., 2023).

4.2. Conclusions

Towards the goal of building more accessible and human-centric legal AI, a high-quality novel dataset containing question-answer pairs was produced and publically released,

[4] `https://txt.cohere.com/how-llms-can-boost-legal-productivity-with-accuracy-and-privacy/`

[5] This was roughly calculated based on the "hundreds of billions" of pages that Google indexes (Google, 2023) compared to our dataset of under a thousand samples.



Question

I'm 16 and my car door flew open and slammed another car and I left. What should I do?
I didn't notice any damage some paint of my car where on the other car but I wiped it.

Expert Answer ✓

Regardless of the nature of the impact, if your vehicle is involved in a collision where the other driver flees the scene, **it's considered a hit and run**. From an insurance perspective, parking lot collisions are treated just like any other in that the same rules are applied to determine who caused the collision ...

GPT-3.5 Web Retrieval ✗

In Ontario, ... If the damage is less than $2,000 and does not cause injuries or damage to property other than the vehicles involved, **it does not require police involvement**. In your situation, since there was minor damage involving paint from your car ... **it may not need to be reported** ...

GPT-3.5 Legal Retrieval ✓

Leaving the scene of an accident, even if you are a minor, is **considered a hit-and-run and can result in legal consequences**. It is crucial to immediately report the incident to your parents or legal guardians and contact the police to report what happened...

Web Retrieved Document ✗

Fails to convey nuance in the situation, misleading the LLM's response.

... Anything less than $2,000 does not require police involvement. The previous minimum for reporting damage was $1,000, which has been unchanged since 1998 when it was raised from $700 ...

Law Retrieved Document ✓

Conveys proper nuance to answer the question, instead of misleading.

If ... car was damaged by an accident or a third party, it depends ... So, who pays for car damage? ... However, if a random act of vandalism or damage to your car (i.e., a light pole falls on your car while on their lot) occurs ...

Figure 6. Qualitative example showcasing retrieval from the entire internet and our high-quality source of legal documents. In this case, retrieval from the entire internet provides a less nuanced source of information.

addressing a previous lack of expert-involved structured data. Existing open-sourced and closed-sourced language models were evaluated using an automatic evaluation framework based on the factuality of answers. We found that retrieval from a small set of legal documents can match or outperform the performance of retrieval from the entire internet, despite requiring many orders of magnitude less data.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Mendes, C. C. T., Chen, W., Chaudhary, V., Chopra, P., Giorno, A. D., de Rosa, G., Dixon, M., Eldan, R., Iter, D., Garg, A., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Huynh, J., Javaheripi, M., Jin, X., Kauffmann, P., Karampatziakis, N., Kim, D., Khademi, M., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Liang, C., Liu, W., Lin, E., Lin, Z., Madan, P., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Song, X., Tanaka, M., Wang, X., Ward, R., Wang, G., Witte, P., Wyatt, M., Xu, C., Xu, J., Yadav, S., Yang, F., Yang, Z., Yu, D., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

Aiden, G. Introducing Command R+: A Scalable LLM Built for Business, Apr 2024. URL https://txt.cohere.com/command-r-plus-microsoft-azure/.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Bhambhoria, R., Dahan, S., Li, J., and Zhu, X. Evaluating AI for Law: Bridging the Gap with Open-Source Solutions, 2024.

Blair-Stanek, A., Holzenberger, N., and Durme, B. V. Can GPT-3 Perform Statutory Reasoning?, 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners, 2020.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LEGAL-BERT: The muppets straight out of law school. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.261. URL https://aclanthology.org/2020.findings-emnlp.261.

Chen, H.-T., Xu, F., Arora, S., and Choi, E. Understanding Retrieval Augmentation for Long-Form Question Answering, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling Instruction-Finetuned Language Models, 2022.

Cui, J., Li, Z., Yan, Y., Chen, B., and Yuan, L. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. *ArXiv*, abs/2306.16092, 2023. URL https://api.semanticscholar.org/CorpusID:259274889.

Dahan, S., Bhambhoria, R., Liang, D., and Zhu, X. Lawyers should not trust ai: A call for an open-source legal language model. *Available at SSRN 4587092*, 2023.

Dai, Y., Feng, D., Huang, J., Jia, H., Xie, Q., Zhang, Y., Han, W., Tian, W., and Wang, H. LAiW: A Chinese Legal Large Language Models Benchmark, 2024.

Duan, X., Wang, B., Wang, Z., Ma, W., Cui, Y., Wu, D., Wang, S., Liu, T., Huo, T., Hu, Z., Wang, H., and Liu, Z. *CJRC: A Reliable Human-Annotated Benchmark DataSet for Chinese Judicial Reading Comprehension*, pp. 439–451. Springer International Publishing, 2019. ISBN 9783030323813. doi: 10.1007/978-3-030-32381-3_36. URL http://dx.doi.org/10.1007/978-3-030-32381-3_36.

Eldan, R. and Li, Y. Tinystories: How small can language models be and still speak coherent english?, 2023.

Gainer, R. and Starostin, K. How LLMs Can Boost Legal Productivity (with Accuracy and Privacy), Apr 2024. URL https://cohere.com/blog/how-llms-can-boost-legal-productivity-with-accuracy-and-privacy.

Google. How google search organizes information, 2023. URL https://www.google.com/search/howsearchworks/how-search-works/organizing-information/.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models, 2023.

Gunasekar, S., Zhang, Y., Aneja, J., Cesar, C., Mendes, T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Singh Behl, H., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. Textbooks are all you need, June 2023. URL https://www.microsoft.com/en-us/research/publication/textbooks-are-all-you-need/.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024.

Jiang, C. and Yang, X. Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, pp. 417–421, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701979. doi: 10.1145/3594536.3595170. URL https://doi.org/10.1145/3594536.3595170.

Ke, Z., Kong, W., Li, C., Zhang, M., Mei, Q., and Bendersky, M. Bridging the Preference Gap between Retrievers and LLMs. *arXiv preprint arXiv:2401.06954*, 2024.

Koniaris, M., Galanis, D., Giannini, E., and Tsanakas, P. Evaluation of Automatic Legal Text Summarization Techniques for Greek Case Law. *Information*, 14(4), 2023. ISSN 2078-2489. doi: 10.3390/info14040250. URL https://www.mdpi.com/2078-2489/14/4/250.

Laurer, M., van Atteveldt, W., Casas, A., and Welbers, K. Building Efficient Universal Classifiers with Natural Language Inference, December 2023. URL http://arxiv.org/abs/2312.17543. arXiv:2312.17543 [cs].

Li, J., Bhambhoria, R., and Zhu, X. Parameter-efficient legal domain adaptation. In Aletras, N., Chalkidis, I., Barrett, L., Goanță, C., and Preoțiuc-Pietro, D. (eds.), *Proceedings of the Natural Legal Language Processing Workshop 2022*, pp. 119–129, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nllp-1.10. URL https://aclanthology.org/2022.nllp-1.10.

Li, Y., Bubeck, S., Eldan, R., Giorno, A. D., Gunasekar, S., and Lee, Y. T. Textbooks are all you need ii: phi-1.5 technical report. September 2023. URL https://www.microsoft.com/en-us/research/publication/textbooks-are-all-you-need-ii-phi-1-5-technical-report/.

Lin, X. V., Chen, X., Chen, M., Shi, W., Lomeli, M., James, R., Rodriguez, P., Kahn, J., Szilvasy, G., Lewis, M., et al. RA-DIT: Retrieval-Augmented Dual Instruction Tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

Lin, Y.-T. and Chen, Y.-N. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In Chen, Y.-N. and Rastogi, A. (eds.), *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pp. 47–58, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.nlp4convai-1.5. URL https://aclanthology.org/2023.nlp4convai-1.5.

Ma, X., Gong, Y., He, P., Zhao, H., and Duan, N. Query rewriting in retrieval-augmented large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5303–5315, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.322. URL https://aclanthology.org/2023.emnlp-main.322.

Oh, J., Kim, E., Cha, I., and Oh, A. The Generative AI Paradox on Evaluation: What It Can Solve, It May Not Evaluate, 2024.

Savelka, J. Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, pp. 447–451, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701979. doi: 10.1145/3594536.3595161. URL https://doi.org/10.1145/3594536.3595161.

Sun, Z., Wang, X., Tay, Y., Yang, Y., and Zhou, D. Recitation-Augmented Language Models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-cqvvvb-NkI.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., Choquette-Choo, C. A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.-C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.-B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L. L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S. L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., and Kenealy, K. Gemma: Open models based on gemini research and technology, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023.

Wang, C., Cheng, S., Guo, Q., Yue, Y., Ding, B., Xu, Z., Wang, Y., Hu, X., Zhang, Z., and Zhang, Y. Evaluating Open-QA Evaluation, 2023.

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. C-pack: Packaged resources to advance general chinese embedding, 2023.

Xu, H. and Ashley, K. Argumentative Segmentation Enhancement for Legal Summarization, 2023.

Yao, H., Rashidian, S., Dong, X., Duanmu, H., Rosenthal, R. N., and Wang, F. Detection of suicidality among opioid users on reddit: machine learning–based approach. *Journal of medical internet research*, 22(11):e15293, 2020.

Yu, F., Quartey, L., and Schilder, F. Legal Prompting: Teaching a Language Model to Think Like a Lawyer, 2022.

Yu, F., Quartey, L., and Schilder, F. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13582–13596, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.858. URL https://aclanthology.org/2023.findings-acl.858.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

Zheng, L., Guha, N., Anderson, B. R., Henderson, P., and Ho, D. E. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pp. 159–168, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385268. doi: 10.1145/3462757.3466088. URL https://doi.org/10.1145/3462757.3466088.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. How does NLP benefit legal system: A summary of legal artificial intelligence. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5218–5230, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.466. URL https://aclanthology.org/2020.acl-main.466.