Answering legal questions from laymen in German civil law system

Marius Büttner[1] and **Ivan Habernal**[2]
Trustworthy Human Language Technologies
[1] Department of Computer Science, Technical University of Darmstadt
[2] Department of Computer Science, Paderborn University
ivan.habernal@uni-paderborn.de
www.trusthlt.org

Abstract

What is preventing us from building a NLP system that could help real people in real situations, for instance when they need legal advice but don't understand law? This question is trickier than one might think, because legal systems vary from country to country, so do the law books, availability of data, and incomprehensibility of legalese. In this paper we focus Germany (which employs the civil-law system where, roughly speaking, interpretation of law codes dominates over precedence) and lay a foundational work to address the laymen's legal question answering empirically. We create GerLayQA, a new dataset comprising of 21k laymen's legal questions paired with answers from lawyers and grounded to concrete law book paragraphs. We experiment with a variety of retrieval and answer generation models and provide an in-depth analysis of limitations, which helps us to provide first empirical answers to the question above.

1 Introduction

As the legal system defines one of the fundamental pillar of democracy, it should be easily accessible for any member of society, regardless of their social background or education. A recent survey of comprehensibility of legal texts in Germany[1] revealed that although searching the internet was the primary choice for 74% of the respondents, most consulted a lawyer afterward claiming that online resources were not helpful enough. Moreover, 31% stated they avoided consulting law books because they do not understand them. Unfortunately, unlike in case-law systems, in Germany's civil-law system law books and their interpretation are the only source of "truth" in legal matters. As a result, individuals must trade off the urgency of their legal problem with the costs of consulting a lawyer, which favors those with more financial resources.

Leveraging NLP tools to address legal question answering has been an active research topic. However, existing works have been focusing on questions asked by *experts*, such as lawyers or legal scholars (Vold and Conrad, 2021; Zheng et al., 2021; Charalabidis et al., 2019). To the best of our knowledge, very few considered the perspective of a *layman*, that is an ordinary person without any legal expertise and, most importantly, without skills in understanding *legalese*, the legal jargon (Butt, 2012).

To fill this research gap, we asked the following research questions. First, how can we best setup empirical research in the domain of legal QA by laymen in a civil-law system? In particular, how can we create a large dataset that (a) contains laymen's questions that are (b) answered by expert lawyers and also (c) grounded in existing law books? Second, to which extent the current transformer-based retrieval models and text-generation models are able to tackle the task? Third, and most importantly, what are the fundamental challenges of this task preventing success of the current state-of-the-art approaches?

This paper presents a ground work for addressing these research questions. We present GerLayQA, a new dataset consisting of 21,538 actual examples for legal German layperson questions accompanied by valid lawyer answers grounded to law books. We then benchmark a variety of pre-trained and/or fine-tuned large language models and semantic retrieval systems. We also conduct in-depth quantitative and qualitative analyses of the results to show the current limitations and where further research is necessary. All datasets, source codes and models are publicly available at https://github.com/trusthlt/eacl24-german-legal-questions.

[1] https://www.bundesregierung.de/breg-de/themen/recht-verstaendlich-machen-1735478

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2015–2027
March 17-22, 2024 ©2024 Association for Computational Linguistics

2 Related work

There has been a growing interest in using NLP to answer legal questions across diverse languages and legal domains. To this end, Kien et al. (2020) conducted a study on the Vietnamese legal system, using a semantic similarity retrieval mechanism to retrieve relevant legal paragraphs in response to questions. Hong et al. (2021) concluded that both extractive and abstractive question answering are still largely unexplored in legal texts. Dale (2019) provided a short survey of services providing legal aid (in English) by navigating users to fill out a predefined form or using a chat-like interface.

In the German legal domain, Hoppe et al. (2021) built one of the first QA datasets by asking lawyers to manually formulate questions for various case law documents. They then further compared the performance of sparse and dense methods for information retrieval and found that the pre-trained BERT model they used could not outperform the sparse retrieval methods. Hoppe et al. (2022) extended their research based on their previously created dataset and focused on developing a system to answer questions by retrieving sub-sections of relevant case law documents for a given query.

Our paper differs from these related works significantly. Firstly, we refer to legal paragraphs (actual sections in law books) instead of legal case documents (e.g., court judgments). Secondly, the existing systems's replies only refer to some significant subsections of documents in their dataset, while we aim to provide an easily understandable answer. Lastly, while Hoppe et al. (2021, 2022) relied on a dataset they manually crafted together with legal experts, we use real-world examples instead.

3 Introducing the GerLayQA dataset

To the best of our knowledge, no dataset that contains real-life examples of legal questions in everyday language, with answers provided by legal experts and, ideally, references to relevant law paragraphs exist, at least not in German as general availability of various legal-related datasets is sparse to non-existent. Therefore, we created such a dataset ourselves by utilizing QA pairs from a German legal online forum `frage-einen-anwalt.de` where laypersons pose their queries and, for a small fee, receive answers from legal experts.

3.1 Quality measures and filtering

We wrote a custom web scraper written in Python. Initially, we extracted more than 180,000 data points by applying Regex-based techniques on the crawled HTML. To ensure that we used only high-quality samples from the raw data, we filtered the dataset as follows:

1. The lawyer's response had to contain references to legal paragraphs to ensure a concrete legal foundation for their answer.
2. The questioner should have rated the lawyer's answer with a rating of three or more out of five stars, indicating its helpfulness and understandability.
3. We limited the questions to 500 tokens to remove several outliers.
4. Since the data we obtained spanned from 2004 to 2023, we excluded answers that referred to outdated or modified laws.

After we applied the above filtering steps, our dataset comprised 43,612 samples. Each sample comprises of the following attributes, as further exemplified in Figure 1:

- **Layman's question** The question the layperson poses in their everyday language.
- **Lawyer's answer** The corresponding response from the lawyer in layperson-understandable language.
- **Relevant paragraphs** A set of specific citations to the German legal codes the lawyer referenced in their response.

3.2 Raw dataset statistics

First, we determine the data quality by evaluating the ratings of registered lawyers and answers. An analysis (see Figure 7 in Appendix B) reveals that most answers have a perfect score of 5.0/5.0, with an overall average of 4.7/5.0. Both metrics show that the scraped data is an excellent foundation for the dataset. We can improve its quality even more by excluding any QA pairs or lawyers rated lower than 3.0, which results only in a minor loss of around 100 data points.

Second, the average problem description has a length of 180 tokens (see Figure 8 in Appendix B). Therefore, we can apply a maximum limit of 500 tokens to remove outliers while including most data points.

Third, we used a simple heuristics to evaluate the complexity of the user's problem description.

Layman's question: New car order without delivery date. It is about the binding order of a new car and the desire to cancel, since no info [...] available.
[...] "Due to the current supply situation, all orders are confirmed WITHOUT delivery date and non-binding subject to production[...] I cancel the contract immediately.
What is the legal situation? [...]

Lawyer's answer: Dear questioner, the passage: "Due to the current delivery situation, all orders are confirmed WITHOUT a delivery date and without obligation subject to production". is ineffective as a general business condition according to § 308 No. 1 BGB.
However, you should set the seller a deadline of two weeks in accordance with § 323 para. 1 BGB [...] to fulfill the contract.
After fruitless expiry of the deadline, you can withdraw from the contract. [...]

Relevant Paragraphs: {§308, §323}

Figure 1: Translated example from the new GerLayQA dataset

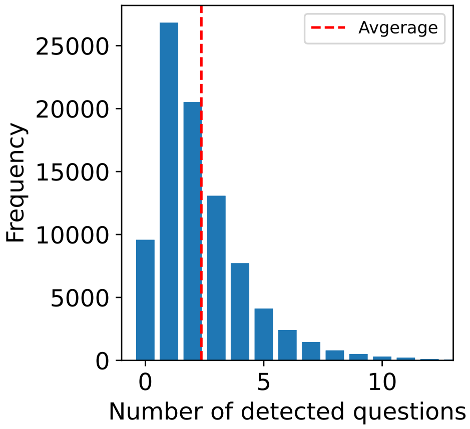

Figure 2: Number of questions raised per post entry

We first examine how an ideal problem description should look. It should start with a detailed description of important background information followed by a precise question for the lawyer. Further, the user's problem description should stay consistent about one legal topic and should not switch contexts. In general, the more questions a user asks within his description, the less of the maximum 500 tokens stand available to provide background information. This makes the problem description more complex for the model and increases the risk of context switching. We detected questions in the description by a set of rules (question mark). Figure 2 reveals that a user poses, on average, 2.4 questions per problem description. We consider data points with more than five questions too complex and remove them later as outliers.

Finally, figure 3 shows the distribution of law books the lawyers cited during their answers. As the most mentioned law books are the BGB, followed by the StGB and the ZPO, and lawyers cite other law books relatively infrequently, we will limit the dataset to these three sources to ensure a reasonable amount of training data for each law. Further details are discussed in Appendix B.

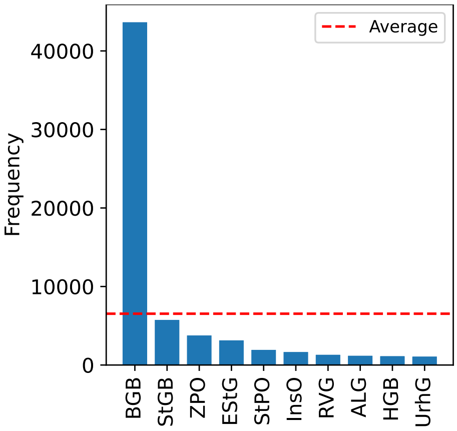

Figure 3: Citations of the top ten German law books

3.3 Final dataset selection

After conducting a thorough analysis of the relevant paragraphs in the dataset, we discovered that the top three most cited law books were the German Civil Code (BGB), the German Criminal Code (StGB), and the German Code of Civil Procedure (ZPO). In response, we developed three separate datasets for each of these law books. However, for our study, we chose to focus on the BGB subset, which consisted of 21,540 samples, since it was the most frequently cited and, therefore, the most relevant to German society. We left the ZPO and StGB subsets aside for future research. We call our final dataset **GerLayQA**—an abbreviation representing laymen legal question answering in German. The

	Train	Val	Test
BGB data points	17,230	2,154	2,154
StGB data points	1,256	157	157
ZPO data points	1,077	135	135

Table 1: Number of data points for the sets after split. Only BGB is currently part of the GerLayQA corpus, with StBG and ZPO left for future work.

GerLayQA dataset is split into train, validation, and test sets (70/15/15), as summarized in Table 1.

4 Experiments

Our approach to answering laymen's legal questions mimics the way humans would solve a legal issue. It consists of a two step approach as shown in Figure 4: document retrieval, which aims to find the most relevant laws to a given question text, and Answer Generation, which should generate an easily understandable answer to the layperson's question.

4.1 Document retrieval

Our first experiment aimed to investigate the effectiveness of existing models in retrieving relevant paragraphs written in legalese to the layperson's question in everyday language. To achieve this, we created embeddings for all the paragraphs in the BGB law book and compared them to the user's question. We then selected the ten paragraphs with the highest cosine similarity score and defined them as the most semantically relevant to the query. We carried out this step on the train set of our BGB dataset using the 'Question text' and the 'Relevant Paragraphs' features of each data point.

To generate the above-mentioned embeddings, we chose several baseline models compatible with the Hugging Face sentence-transformers library. This made producing and comparing their embeddings easy since the sentence-transformers library builds upon SBert's bi-encoder architecture (see Figure 5). After exploring various models, we identified two with notable potential: The `PM-AI/german`,[2] specifically trained for the document retrieval task, and the `bert-base-german-cased` model, which has its foundation in legal texts. Additionally, we included OpenAI's `text-embedding-ada-002` model in our search, drawn to its extensive and varied training dataset.

4.1.1 Evaluation measures

For evaluating our model performances, we applied the standard metrics for the document retrieval task: Precision, Recall, F1 Score, alongside the advanced ranking metrics Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP). These metrics collectively offer comprehensive insights into the models' retrieval accuracy, their ability to capture all relevant paragraphs, and their effectiveness in ranking relevant documents.

In order to the models within our task-specific performance range, we defined the following borders for our evaluation dataset and a $\mathrm{top}_k = 10$ retrieval.

Random baseline To represent the minimum expected performance, we randomly selected ten paragraphs from our document collection. We compared them to each data point's gold standard of relevant paragraphs.

Oracle upper bound We simulate almost perfect performance by allowing the model to make an error on one of the expected paragraphs. We model this by randomly replacing the last item from each question's gold standard sorted list or relevant paragraphs. If the data point only had one relevant paragraph, we simply kept it unchanged. Our final lists for calculating the oracle upper bound therefore contains n elements starting with the relevant paragraphs of each data point.

4.1.2 Document retrieval results

After analyzing the performance of the selected models on our test dataset, we found that all the baselines showed moderate results, as shown in Table 2. While the `text-embedding-ada-002` model performed the best, followed by the `PM-AI/german` model, the legally pre-trained `bert-base-german-cased` model only managed to achieve scores slightly better than the random baseline.

Since the sentence-transformer library provides a simple way to fine-tune its compatible models, we chose to use the `PM-AI/german` model and our BGB train dataset to do so. We experimented with two different loss functions (`CosineSimilarityLoss` and `TripletLoss`) to optimize our performance but unfortunately did not achieve the desired results.

[2] `PM-AI/bi-encoder_msmarco_bert-base_german` is the full identifier on HuggingFace

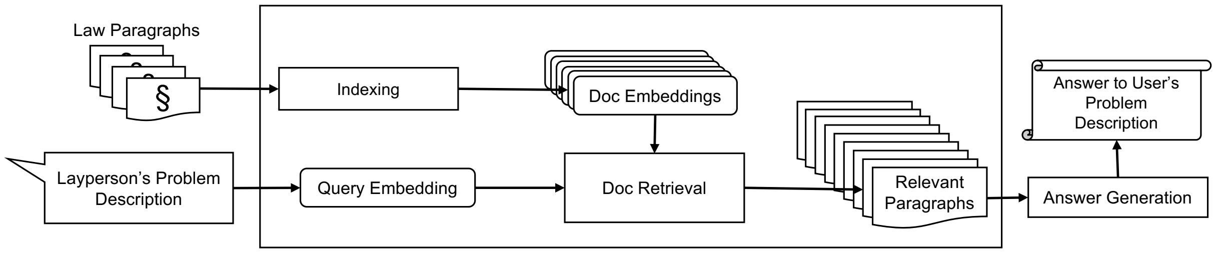

Figure 4: Architecture of the QA pipeline

Model	Prec.	Rec.	F_1	MRR	MAP
Random Baseline	0.001	0.004	0.001	0.002	0.001
Oracle upper bound	0.131	0.831	0.215	1.000	0.831
OpenAI					
— text-embedding-ada-002	0.033	0.226	0.055	0.146	0.108
Sentence transformers (HF)					
— PM-AI/German	0.026	0.176	0.044	0.117	0.089
— T5-base	0.006	0.039	0.011	0.025	0.015
— bert-base-german-cased	0.005	0.035	0.009	0.022	0.015

Table 2: Results for $\text{top}_k = 10$ document retrieval; HF = Hugging Face

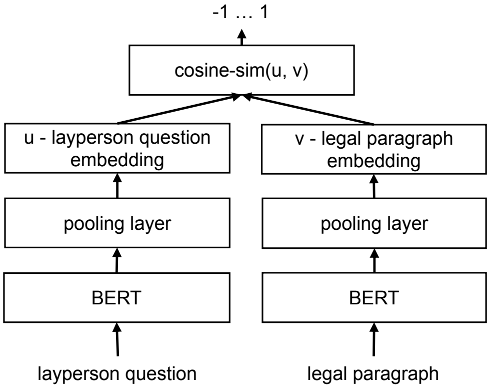

Figure 5: Bi-Encoder setup

4.2 Answer generation

Moving to the second part of our pipeline, we want to examine how effective NLP models can provide an answer to legal queries. We expected the model to generate answers in natural fluent language, containing all essential details from the paragraphs.

4.2.1 GPT-3.5-turbo with legal paragraphs

For this step, we rely once more on the evaluation BGB dataset, this time with the following included features:

- The question posed by the layperson
- The relevant 'gold' paragraphs that the lawyer cites in his answer
- The 'gold answer' given by a lawyer to the layperson's question

As our previous document retrieval Task showed a relatively moderate performance, we decided to use the gold paragraphs instead of the retrieved results from earlier. With that, we aim to unlock the full potential of the Answer Generation model by working with a reliable source of references.

As we already used a model of OpenAI and were therefore familiar with their easy-to-use API, we decided to go with their GPT-3.5-turbo model. In order to generate an answer for each of our data points, we queried the model with the following prompt and stored the result in the new feature 'Generated Answer' in our data set:

> *"Answer the following question: {layman's question} Based on these legal paragraphs: {set of legal paragraphs}"*
>
> *Original query: "Beantworte folgenden Frage: {layman's question} Auf Grundlage dieser Gesetzestexte: {set of legal paragraphs}."*

4.2.2 GPT-3.5-turbo turbo without legal paragraphs

After exploring the use of legal paragraphs to guide the model's responses, we shifted focus to understand how the GPT-3.5-turbo model performs when relevant legal references are omitted. This phase aimed to simulate a scenario where laypersons utilize the model without prior access to specific legal documentation. Here, we presented only the layperson's question to the model, excluding any legal paragraphs used in previous experiments. The prompt for this test was adpated to:

> *"Answer the following question: {layman's question}."*
>
> *Original query: "Beantworte folgenden Frage: {layman's question}."*

4.2.3 Evaluation measures

We evaluate our model's Answer Generation performance using the ROUGE score and the BERTScore. While ROUGE compares the lexical similarity between the generated and gold answers by counting overlapping n-grams (Lin, 2004), we rely on BERTScore to evaluate the semantical matching between candidates and references through cosine similarity using pre-trained BERT models (Zhang et al., 2020). With that, it provides a more fine-grained evaluation than ROUGE by considering the contextual embeddings of words.

In order to evaluate the performance of our model in a task-specific range, we follow a similar approach to our document retrieval task.

Lower baseline To determine the minimum expected performance of our model, we compare the relevant paragraphs of each data point to the lawyer's answer using the aforementioned metrics. This simulates a real-world scenario where a layperson reads legal sources to obtain information about their issue.

Oracle upper bound For our Oracle upper border, we envision an ideal scenario where the generated answer includes all the essential legal details the lawyer included in his gold answer, albeit potentially formulated in a different style or wording.

Such an alternative-generated answer would be the equivalent of another lawyer answering the question while using a different explanation of the same legal advice from the lawyer within our gold answer.

Due to our limited legal knowledge and resources, we could neither rephrase the sentences ourselves nor consult with lawyers to do so. Hence, we instead relied on the capabilities of the GPT-3.5-turbo model for rephrasing tasks and prompted the model to create a rephrased version of each gold answer with the following query:

> *"Rewrite this text, but keep all the information! {layperson's question}."* [3]

After generating the rephrased answers and manually verifying a subset of them for quality, we applied selected metrics to compare the gold and rephrased answers and establish an Oracle upper border comparison.

4.2.4 Quantitative baseline evaluation: analysis of generated metrics

By applying the ROUGE and BERT scores to the generated and gold answer, we obtained the model performance, as displayed in Table 3.

After analyzing the ROUGE scores, it is evident that the GPT-3.5-turbo baseline model generates answers that contain overlapping n-grams and longer sequences compared to the gold answer of our lawyer. Therefore, it performs significantly better than our Lower baseline, indicating that its generated answers are easier to understand and their relevant information is more accessible for laypeople than the original texts in legalese.

However, the model still lags significantly behind the Oracle upper border, which suggests that there is still room for improvement, and it cannot compete with human-generated answers by legal experts.

When it comes to BERTScore, the model equally outperforms the lower baseline. It is noteworthy that the Random baseline is relatively high, which is unexpected considering that it compares the paragraphs in legalese and the gold answer in a more natural language. We can attribute this to the limited capability of the underlying model to differentiate and understand the legal nuances in German texts while calculating the BERT score.

Nevertheless, the BERTScore performance indicates that the model includes many key concepts in their generated answers that are likewise present in lawyers' answers. Comparing the GPT-3.5-turbo model's performances (excluding and including additional relevant laws paragraphs, respectively) shows a modest enhancement when law references are provided. This slight improvement highlights that the process of sourcing relevant legal paragraphs may not be essential for laypersons seeking initial legal advice. It suggests a more accessible

[3] Original query: "Schreibe diesen Text um, aber behalte alle Informationen! {lawyers's answer}."

Metric	Lower baseline	GPT-3.5-turbo without legal paragraphs	GPT-3.5-turbo with legal paragraphs	Upper bound
ROUGE-1	0.1463	0.2512	0.2910	0.4613
ROUGE-2	0.0108	0.0430	0.0646	0.2812
ROUGE-L	0.0711	0.1078	0.1244	0.3747
BERTScore	0.6185	0.6364	0.6550	0.7478

Table 3: Results of GPT-3.5-turbo on answer generation

approach for the general public, indicating that satisfactory legal guidance can be obtained even without the intricate step of navigating through legal texts.

Before moving on to a manual inspection of the generated answers, we can conclude that the model significantly outperforms our Random baseline while falling short of the ideal performance of the Oracle upper border.

5 Analysis and discussion

This section presents a detailed manual look at the results of our two stages. With this, we aim to identify possible challenges the models faced, influencing their performances.

5.1 Document retrieval in-depth analysis

We analyzed the 25 best and 25 worst data points regarding their precision score to compare the characteristics for which the models achieved better or worse performance. We found that the only visible differing aspect was the relationship between the texts' length and the precision of the model.

5.1.1 Length influence on retrieval

In high-precision examples, we noticed that the laypeople's problem descriptions were short and informative, averaging around 140 tokens. The corresponding true positive paragraphs for these questions were also short, averaging around 120 tokens. This indicates that the model performs better on texts of moderate length where necessary information is densely packed and more directly correlated. Additionally, the high-precision problem descriptions often included specific buzzwords that might indicate a closer match with the related legal paragraphs.

In contrast, low-precision examples typically consisted of longer problem descriptions, averaging about 243 tokens, with notably shorter relevant paragraphs, averaging 90 tokens. Examining these problem descriptions further, why this challenges the model: When laypeople describe their legal issues, they often include too many details, confusing the model with what is relevant. While these details may seem important to the speaker, they can distract the model's attention. As a result, the created embeddings focus more on these unimportant details than the central legal issue, resulting in poor context representation.

5.1.2 Semantic relevance of retrieved paragraphs

We stumbled upon an unexpected observation while analyzing the retrieved paragraphs. Interestingly, many of these 'false positive' paragraphs, whether high-precision or low-precision examples, appeared to be somehow contextually related to the problem description, at least from a layperson's perspective.

This led us to question whether some retrieved paragraphs could be relevant to the legal issue, even if the lawyer did not cite them in their gold-standard answer. In such cases, the model's false positives might not be entirely incorrect, but lawyers may have simply not cited them, as they are not the primary legal reference they used. However, to definitively evaluate the actual relevance of each paragraph, we need further insights from legal professionals. Thus, our analysis remains based on the scores from our created dataset.

5.1.3 Embedding space analysis

A deep dive into the vector space of the embeddings further illustrates how good the model's embeddings are. By displaying the embedding's vectors as in Figure 6, we can observe that the model embeds some false positives closer to the problem description than the true positives. This clearly indicates that the model, in its current state, cannot create accurate enough embeddings for retrieval purposes. Applying further fine-tuning or training a model from scratch on such a task could be beneficial to optimize these created embeddings.

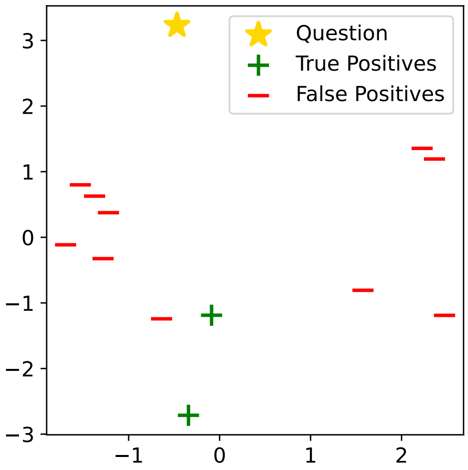

Figure 6: PM-AI/german's true and false postives embedded in relation to question's embedding for document retrieval

5.2 Answer generation analysis

In addition to relying on the above metrics, we manually inspected the generated answers to determine how understandable and valuable they are to non-legal experts. Therefore, we will examine the most striking differences between the model's and the lawyer's answers.

5.2.1 Approach to addressing legal issues

When examining how the lawyers and the AI model address legal matters, a noticeable difference emerges between AI-generated and the lawyers' responses.

The AI model generally rephrases and simplifies the legal clauses, aiming to relate them to the described legal issue. While this cautious approach might serve as a practical first introduction to the layperson's legal matters, it often does not offer a concrete solution path.

On the other hand, lawyers tend to provide more direct responses. In addition to answering the question, they often suggest additional assistance or outline concrete next steps. This not only offers clear guidance for the next course of action but also displays a solid understanding of legal expertise and knowledge.

5.2.2 Language difficulty of the answers

The second aspect we noticed for specific answers was a difference in language complexity.

Starting with the lawyers' answers, they frequently integrate direct citations of legal text. While this provides a strong substantiation for their advice, it also may reduce the accessibility for laypersons as the legal terminology is more difficult to understand.

Looking at the generated answer, we can see the exact opposite. By using simplified language, the reply is easily understandable by laypersons, but by leaving out relevant citations or references to the legal paragraphs, the answer also feels less legally accurate.

5.2.3 Insufficient question's context or details

Going forward, we observed a significant difference in the approaches adopted by the two parties while dealing with queries with limited or incomplete input information.

Analyzing the response generated by the AI model, we noticed that the model leans heavily on the input data and often overlooks the possibility that additional information may exist that could be crucial for nuanced legal advice. This can be challenging for a layperson who may not have provided all the relevant case details.

On the other hand, the lawyers, with their legal expertise, can identify such matters where potentially missing information would significantly affect the legal outcome and highlight them while answering the question.

In extreme cases, if the question texts or the relevant paragraphs provide insufficient information, the GPT-3.5-turbo model acknowledges that it cannot answer due to the lack of data. In comparison, the legal experts try to assist the questioner with their answers to close that gap. They do that by identifying additional sources of information or highlighting the use of which information could be helpful. Although the legal experts cannot provide a definite answer, their approach is much more helpful for the users, as they can refine the formulation of their legal issue for the next time.

5.2.4 Answer quality for laypersons

Apart from the special cases described above, the model was generally able to generate an answer that aligned with the legal core statements of the lawyer, at least from a layperson's perspective. For a well-formulated question, the model extracted the relevant information from the relevant legal paragraphs and presented it in an easy-to-understand language to the questioner.

However, we must emphasize that we cannot verify the validity of the model's answers due to the lack of legal expertise. Instead, we can only state that, according to our understanding as laypersons, the content of the generated answers often matches the essence of the lawyer's answers.

6 Conclusion

This paper introduced GerLayQA, the first German dataset designed for laypeople seeking legal advice. Our dataset comprises real-life examples of questions asked by laypeople, lawyers' corresponding answers, and relevant legal paragraphs. By including all relevant aspects when working on legal cases, we have created a comprehensive database for our and future research that aims to assist laypeople in seeking legal guidance.

We experimented with a two-step QA pipeline, similar to the workflow used by lawyers. We found that all tested models delivered only moderate performances. For the most hindering aspects of the models' performances, we identified their difficulties in understanding German legal texts. Since the models were not trained on legalese, creating proper semantic embeddings for this formal language is challenging. As a result, especially for document retrieval, using such embeddings to compare the semantic meanings for paragraphs in legalese and questions in everyday language only produces moderate retrieval results. Furthermore, the models struggle in essential tasks to grasp legal nuances and understand legal correlations due to insufficient training when providing a legal answer.

Future work Training a bespoke model for German laymen's and expert legal text analysis, similar to the LEGAL-BERT (Chalkidis et al., 2020), might improve the accuracy and efficiency of laymen legal QA. Further, including more legal expertise in evaluation can be highly beneficial. It would allow us to assess the model's outputs with a legal background rather than just comparing them to the provided gold standard using various automatic metrics. For document retrieval, this legal knowledge would provide us with a more accurate means of determining the relevance of each retrieved paragraph to the query instead of relying on binary labels. Similarly, for answer generation, legal experts could aid in validating the legal soundness and binding nature of the answers rather than solely relying on statistical metrics like ROUGE and BERT-Score.

Limitations

While our research provides new valuable insights into the unexplored German legal domain, it is at the same time limited in scope and requires careful interpretation.

Privacy considerations For our data extraction process, we ensured that we followed ethical privacy and only extracted information from publicly accessible sections of the legal online forum. As the platform already provided anonymity for questioners, we did not need to take additional steps to protect their identities.

BGB and German legal domain Our exploration mainly focuses on the laws of Germany's legal system. Therefore, all models and the dataset itself may not adapt to other legal landscapes, even for other German-speaking countries.

Moreover, even within Germany, our study mainly worked with the BGB. In reality, the German legal environment includes many more law books. Consequently, it is essential to take the whole landscape into account before considering such a system to be able to give legally binding advice. By focusing on a subset, the system will inadvertently miss crucial legal aspects and provide incorrect legal advice.

Dataset Limitations Our dataset, based on real-life queries from laypersons, is complex and presents a challenge when it comes to filtering for semantically sound questions. Despite our efforts to remove poorly rated QA pairs from the dataset, we still encountered queries that lacked sufficient information to provide an accurate answer. To address this issue, we suggest that legal experts manually filter the dataset to remove these unhelpful queries. Legal experts are better suited for this task as they can identify questions with inadequate information.

Moreover, the complexity of the questions can lead to varying interpretations by legal professionals, resulting in different gold standard answers. To tackle the existing issue and enhance the accuracy of our dataset, we suggest engaging a secondary lawyer to support and verify the gold standard responses. This would increase the trustworthiness of the gold answers.

Evaluation limitations Our evaluations have revealed that our lack of legal expertise limits our model's performance. We believe that having legal experts on our team would provide us with valuable insights into the model's actual performance. For the document retrieval step, experts could assess the relevance of retrieved paragraphs to our query, which would enhance our model's performance. From our dataset's setup, we could only use the binary labels approach to classify documents as relevant. Similarly, within our answer generation step, we could define more precisely how legally suitable the generated answer is to the provided question. By incorporating legal experts in both steps, we can train our models with more accurate data, improving our overall performance.

Ethical considerations Besides technical aspects, it is important to consider ethical considerations when providing legal advice through NLP models. Legal advice carries significant responsibility due to the severe consequences of misguided counsel. Therefore, it is essential to raise awareness amongst users regarding whether they have received advice from an NLP model or a certified legal lawyer.

In conclusion, while current NLP models can provide additional insights into a layperson's legal questions, they cannot replace the role of a human lawyer in delivering a legally valid response. As these models will improve in the near future, it is essential to address ethical considerations in this field. It is crucial to ensure transparency and raise awareness among users that they receive legal advice from an NLP model, not a certified lawyer. As the answers generated by the model are likely to become closer to those of human experts, it is vital to prevent society from lawsuits and reliance on non-binding legal advice.

References

Peter Butt. 2012. Legalese versus plain language. *Amicus Curiae*, 2001(35).

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Yannis Charalabidis, Michalis Avgerinos Loutsaris, Shefali Virkar, Charalampos Alexopoulos, Anna-Sophie Novak, and Zoi Lachana. 2019. Use Case Scenarios on Legal Text Mining. In *Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance*, pages 364–373, Melbourne, Australia. ACM.

Robert Dale. 2019. Law and Word Order: NLP in Legal Tech. *Natural Language Engineering*, 25(1):211–217.

Jenny Hong, Catalin Voss, and Christopher Manning. 2021. Challenges for Information Extraction from Dialogue in Criminal Law. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 71–81, Online. Association for Computational Linguistics.

Christoph Hoppe, Nico Migenda, David Pelkmann, Daniel Hötte, and Wolfram Schenck. 2022. Collaborative system for question answering in german case law documents. In Luis M. Camarinha-Matos, Angel Ortiz, Xavier Boucher, and A. Luís Osório, editors, *Collaborative Networks in Digitalization and Society 5.0*, volume 662 of *IFIP Advances in Information and Communication Technology*, pages 303–312. Springer International Publishing, Cham.

Christoph Hoppe, David Pelkmann, Nico Migenda, Daniel Hotte, and Wolfram Schenck. 2021. Towards intelligent legal advisors for document retrieval and question-answering in german legal documents. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 29–32. IEEE.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998, Stroudsburg, PA, USA. International Committee on Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Andrew Vold and Jack G Conrad. 2021. Using Transformers to Improve Answer Retrieval for Legal Questions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 245–249, Online. ACM.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168, Online. ACM.

A Original unabridged example in German

Layman's question Neuwagenbestellung ohne Liefertermin. Es geht um die verbindliche Bestellung eines Neuwagens und dem Wunsch der Stornierung, da keinerlei Infos zum Vorgang

seitens des Herstellers, auch auf mehrmaliger Nachfrage hin, nicht vorliegen.
Problem beim Passus im Kaufvertrag:
"Verbindliche Bestellung zu den nachfolgenden Bedingungen und unter Einbeziehung der beigefügten Neuwagen-Verkaufsbedingungen ("NWVB") folgendes Kraftfahrzeug in Serien-/Sonderausführung..."
"Aufgrund der aktuellen Liefersituation werden alle Bestellungen OHNE Liefertermin und unverbindlich vorbehaltlich einer Produktion bestätigt."
Heißt für mich lapidar: Auto kann aber muss nicht gebaut werden und wenn, dann ist unbekannt wann geliefert wird.
Ist das so rechtens?
Bestellung Neuwagen 20.04.2022. Seitdem keine Infos zu Produktionsstatus, Sachstand, Bestellvorgang als solcher. Vom Verkäufer wurde kein Liefertermin, auch nicht unverbindlich genannt, da es Vorgabe vom Hersteller sei, keine Angaben zu machen. Vom Verkäufer wurde mir zwar eingeräumt, bei Nichtlieferung ab 12 Monate seit Bestellung, den Vertrag kostenlos stornieren zu können, allerdingsf möchte ich den Vertrag sofort stornieren.
Wie sieht die rechtliche Lage aus? Eine Einigung sollte aufgrund der eingeräumten Frist zeitnah und vorgerichtlich erzielt werden.

Lawyer's answer Sehr geehrte/r Fragesteller/in,
der Passus: „Aufgrund der aktuellen Liefersituation werden alle Bestellungen OHNE Liefertermin und unverbindlich vorbehaltlich einer Produktion bestätigt." ist als allgemeine Geschäftsbedingung gem. § 308 Nr. 1 BGB unwirksam.
Sie sollten dem Verkäufer aber gem. § 323 Abs. 1 BGB noch beweisbar (d. h. schriftlich per Einwurf-Einschreiben) eine Frist von zwei Wochen setzen, um den Vertrag zu erfüllen.
Nach ergebnislosem Ablauf der Frist können Sie vom Vertrag zurücktreten.
Ich hoffe, Ihnen mit diesen Auskünften gedient zu haben und weise darauf hin, dass diese auf Ihren Angaben beruhen. Bereits geringfügige Abweichungen des Sachverhalts können zu einer anderen rechtlichen Bewertung führen.
Nutzen Sie bei Rückfragen gern die kostenlose Nachfragefunktion!
Mit freundlichen Grüßen
PersonXY
Rechtsanwalt

Relevant paragraphs {§308, §323}

B Additional raw data analysis

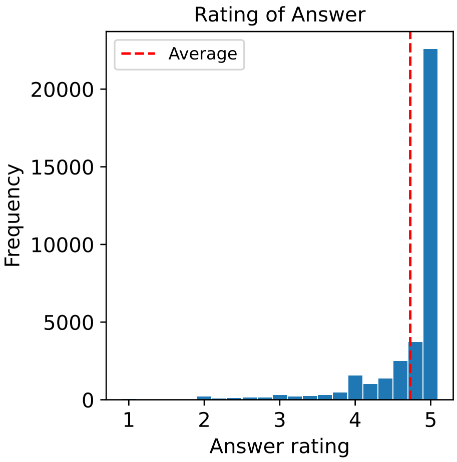

Figure 7: Distribution of answer ratings by the questioner

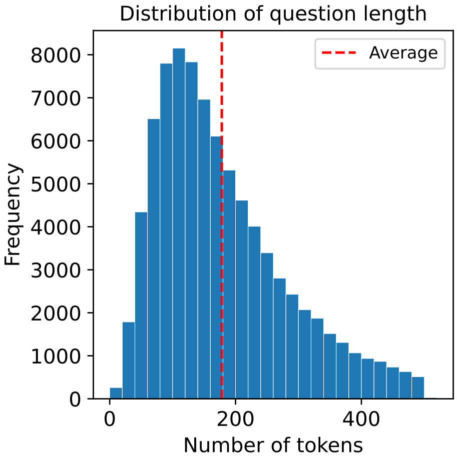

Figure 8: Distribution of question length, limited to a maximum of 500 tokens

Areas of law To inspect what legal directions the scraped data involves, we examine the legal categories tagged to the QA pairs and display them in Figure 3. After grouping all questions by their category tag, the top five categories are:

- Tenancy law, condominium law (Mietrecht, Wohnungseigentumsrecht)
- Labor law (Arbeitsrecht)
- Family law (Familienrecht)
- Contract law (Vertragsrecht)
- Inheritance law (Erbrecht)

The distribution of these categories shows that the platform is primarily used by citizens seeking help for legal problems in their private lives.

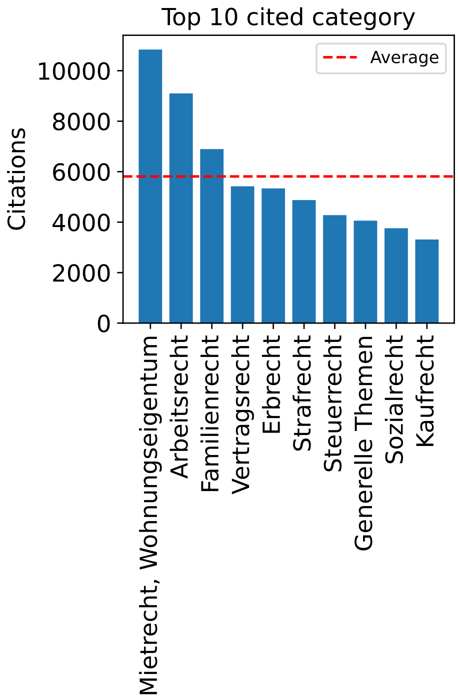

Figure 9: Distribution of top ten question-categories

Platform usage trends Many citizens seek alternative ways to receive legal help instead of visiting a lawyer. We can support this hypothesis by examining the platform's usage over the last few years. As Figure 10 shows, the demand is high based on the number of questions yearly, but a recent decline has occurred. One explanation for this decline is the introduced 'premium feature' on frag-einen-anwalt.de. If a user sets a price of over 35 € on their question, it becomes inaccessible to the public and is therefore not included in this metric.

Price tags To conclude our analysis, we look at the amount of money users offered the lawyers for solving their legal problems. As this property was only accessible for a small subset of the data, it may not fully represent the scraped examples. With a range of 25 - 400 € and an average of 64€, the prices are relatively low compared to the average hourly wage of a lawyer, between 142 - 252 €.[4] On the one hand, this shows that users use this website primarily for more minor legal matters that are not critical enough to consolidate a lawyer. On the other hand, we can consider this a clear indicator of how much money users are willing to pay to receive legal help in their personal life.

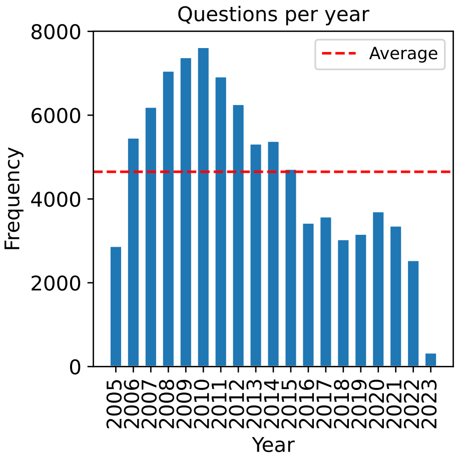

Figure 10: Usage of 'frag-einen-anwalt.de' between 2004 and 2023

C Data point examples for the answer generation analysis (Section 5.2)

C.1 Example for approach to addressing legal issues

Question text "[...] The only bed they offered me appealed to me. The price [...] was [...] 4300€ without mattress. [...] At home, I wanted to know more about the manufacturer and could hardly believe my eyes. The totally same bed catalog price for 1895 € and free shipping. My question now, can I withdraw from the purchase under these machinations?"

Gold answer "[...] here the way via § 138 BGB is certainly well feasible: [...] If you like, I can already write to the seller on your behalf and ask for the money back. [...]"

Generated answer "According to the information available, this could possibly be an immoral legal transaction. [...] It is recommended that you contact a lawyer [...]"

C.2 Example for language difficulty of answers

Question text "Hello, I am the owner of a small PV system (9.9 kW) on my house. The installer refuses to give me the 'login data' for the master data register. This makes it impossible for me to make further entries in the future. Due to the large

[4] https://www.brak.de/presse/zahlen-und-statistiken/star/star-2020/abrechnung-ueber-zeithonorare/

number of applications for the PV system, it is possible that I have authorized him to make the entry/opening. But I do not know. Am I entitled to have the login data handed over? Should I give him a deadline for handing them over? Thank you very much"

Gold answer "[...] see § 273 BGB in this regard: If the debtor has a due claim against the creditor arising from the same legal relationship on which his obligation is based, he may, unless otherwise stated in the obligation, refuse to perform until the performance due to him is effected (right of retention). [...]"

Generated answer "[...] The right of retention can only be asserted if the installer has a claim against you that is due and arises from the same legal relationship on which your obligation is based [...]"

C.3 Example for insufficient question's context or details

Question text "Hello, I am the owner of a small PV system (9.9 kW) on my house. The installer refuses to give me the 'login data' for the master data register. This makes it impossible for me to make further entries in the future. Due to the large number of applications for the PV system, it is possible that I have authorized him to make the entry/opening. But I do not know. Am I entitled to have the login data handed over? Should I give him a deadline for handing them over? Thank you very much"

Gold answer "[...] I cannot judge whether the installer is entitled to a right of retention from another point of view, see § 273 BGB: [...]"

Generated answer "[...] It is not apparent that you have a due claim from the installer that justifies the right of retention. Therefore, the installer cannot refuse to hand over the login data. [...]"

C.4 Example for answer quality for laypersons

Question text "Hello, brief history: Ex 3 months pregnant, separates, finds a new partner 3 weeks later (Jan. 06), [...]. Wanted to see my daughter and got a report for trespassing, I get reported for every little thing. All these allegations are far-fetched and I was able to prove this in the end. [...] Now I receive another letter from this lawyer, accusing me of defamation (i.e. knowingly committing a crime?) and asking me to sign a cease-and-desist declaration. [...] I [...] would like to take [...] action against the lawyer. The [...] What can I do [...]? Thank you in advance"

Gold answer "[...] First of all, I assume that the lawyer is acting on behalf of your former partner (LG); however, this is not yet entirely clear from your comments. However, it is important to know whether the lawyer acted personally (in his own name) or on behalf of your former partner. Because depending on this, the person who is asserting claims for injunctive relief here [...]"

Generated answer "I am an AI model and cannot give legal advice. However, it is advisable that you contact a lawyer for appropriate advice and legal support. A lawyer can help you assess the situation and advise you on the best options for proceeding."