ESG-Consultant: Developing of an ESG Compliance Consulting Tool for Companies Using RAG

Anonymous ACL submission

Abstract

Environmental, Social, and Governance (ESG) compliance has emerged as a crucial benchmark in contemporary corporate assessment. The present work develops and evaluates a software tool called *ESG-Consultant* to facilitate ESG compliance for companies. We also formulate two tasks related to the application of ESG: determining the *Applicability of ESG Law* to companies and *ESG Law Question Answering*. To the best of our knowledge, no previous works tackle these tasks specifically on ESG regulations. For both tasks, we create datasets and develop pipelines based on Retrieval Augmented Generation using LLMs. The results are quite promising, proving that our system can be effectively used to support, but under no circumstances replace practitioners. We hope this work contributes to more effective corporate practices for a sustainable future. The code of our tool and demonstration system are available online[1].

1 Introduction

Recently notable advances have been achieved in Natural Language Processing (NLP) tasks due to the advent of Large Language Models (LLMs) (Patil and Gudivada, 2024; Zhang et al., 2023b). They achieve state-of-the-art performance (OpenAI, 2023) and are highly flexible, to the point they can be used for new tasks with just a few examples, or even without further training while still showing outstanding performance (Brown et al., 2020a; Kojima et al., 2022). One area of knowledge where this technology is increasingly being used is the legal domain (Ganguly et al., 2023; Martinez-Gil, 2023). Legal NLP has been historically challenging due to its highly specific domain language, sometimes referred to as legal "jargon" (Katz et al., 2023). Another limiting factor is that the legal domain is a high-stakes domain, where outcomes and decisions can have a big impact on people and companies. Moreover, legal texts understanding might also require a reasoning step (Shui et al., 2023), which only recently is being observed in modern LLMs (Qiao et al., 2023; Kojima et al., 2022; Wei et al., 2022).

This paper focuses on the Environmental, Social, and Governance (ESG) regulations which have become pivotal benchmarks in contemporary corporate assessment (Li et al., 2021; Gillan et al., 2010). ESG is crucial for sustainable economic development, requiring businesses to positively impact their employees, families, communities, and society by aligning with sustainable development goals. The number of ESG regulations is growing significantly over the past decades, exceeding 2000 globally (ESG, 2023). Many of these new regulations are mandatory, which pushes the companies to assess, understand, and implement ESG regulations using automatic methods.

To the best of our knowledge, there is no published research implementing and evaluating a system using LLMss to assist in ESG compliance. Therefore, in this paper, we develop and evaluate a system to fill the gap, gathering requirements from ESG practitioners. We formulate two NLP tasks related to the application of ESG. The first is the **ESG Law Question Answering task (ESG-QA)**, a vanilla Question Answering task with the texts from the ESG domain. The second is the **ESG Law Applicability task**, a standard classification task, which aims at determining whether a ESG regulation applies to a company with specific parameters.

The outcome of our research is a system called **ESG-Consultant** which aims to support companies and practitioners with Environmental Social, and Governance (ESG) compliance. The implemented prototype of **ESG-Consultant** is deployed, realizing our users' requirements, and manually tested by ESG experts. Additionally, the tool includes

[1] https://anonymous.4open.science/r/ai4esg_experiments-2853/

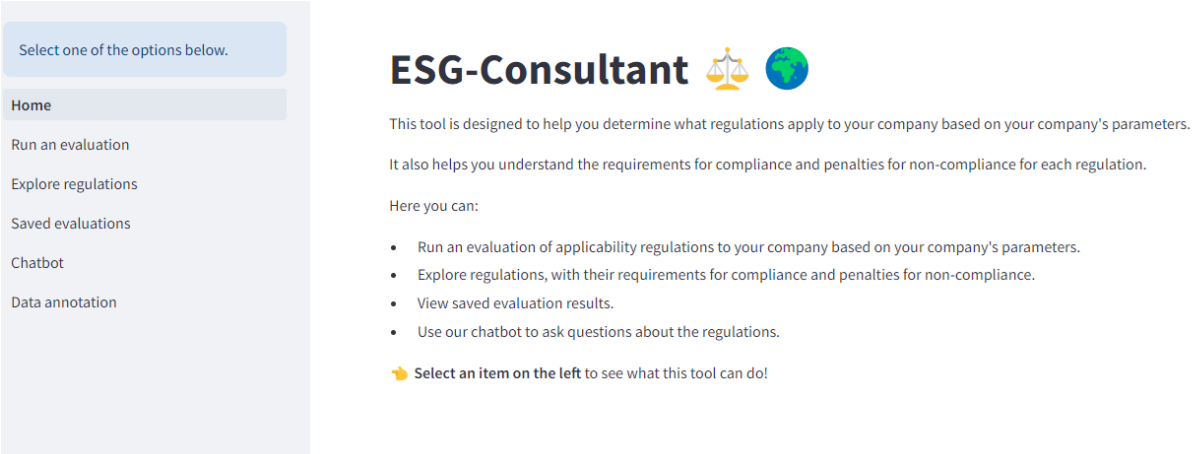

Figure 1: Home screen of the ESG-Consultant tool.

an annotations module, where users can annotate examples for the ESG Law Applicability task for further improvement of the system.

All in all, the main contributions of the paper are as follows:

- We present a demonstration system for assisting law experts with ESG regulations as well as present the methodology of developing such systems.
- We introduce two novel domain-related tasks: ESG Law Applicability and ESG Law Question answering tasks, and create two benchmarks for their evaluation.
- We present the baselines for both tasks based on the Retrieval Augmented Generation (RAG) pipeline (Lewis et al., 2020) using LLMs, achieving high (about 90%) recall results for the ESG-QA task.
- We present an additional annotation module, for creating manual annotations for ESG regulations, which can be used for further improvement of the system.

It is worth noting that our system could be implemented not only for ESG regulations but also for any type of regulatory or legal texts. We use Apache License, Version 2.0, therefore, it can be also re-implemented for other legal tasks.

2 Related Work

In this section, we show the evolution of NLP in the legal domain in general, and the relation between ESG and Artificial Intelligence (AI).

2.1 NLP in the Legal Domain

More and more research on legal tasks now make use of LLMs (Katz et al., 2023; Ganguly et al., 2023). This trend includes methods, applications, and novel datasets for the field (Guha et al., 2024; Martinez-Gil, 2023; Fei et al., 2023; Ganguly et al., 2023). Nay et al. (2024) analyses the impact of different prompting techniques for legal understanding through questions, seeing performance increases from both chain-of-thought prompting (Wei et al., 2024), and few-shot prompting, a form of in-context-learning (Brown et al., 2020b).

Notwithstanding these advancements, the most dangerous risk is the hallucination of LLMs, especially in the high-stakes domains, i.e. law or health. Dahl et al. (2024) claims that LLMs hallucinate between 58% and 82% of the time when used on their own for legal domain tasks. A recent study by Magesh et al. (2024) on RAG tools for the legal domain showed that these tools hallucinate less than LLMs alone.

In our research, we also try to avoid hallucination: in addition to RAG, we force the model to copy the exact excerpts instead of generating text.

2.2 Question Answering in the Legal Domain

For the specific tasks of question answering Martinez-Gil (2023) makes a wide survey of state-of-the-art legal Question Answering (QA) systems using the BERT architecture (Hoppe et al., 2021; Huy et al., 2021; Abualhaija et al., 2022) and knowledge graphs (Dai et al., 2022; Thomas and Sangeetha, 2022). They explain that the main objective of these systems is to assist legal practitioners with the vast amount of legal information existing and the rapid release of new law, which causes the practice of law to be more costly.

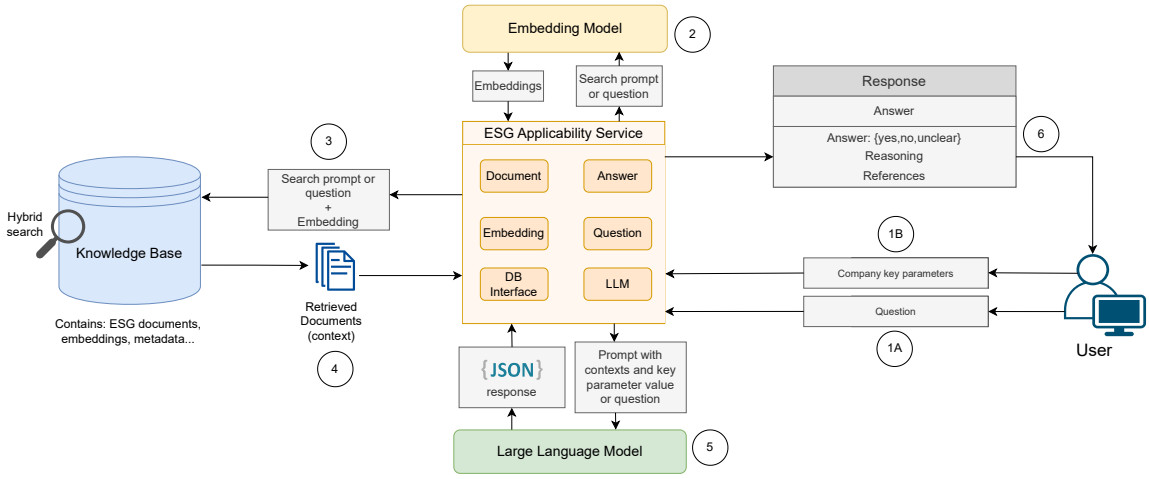

Figure 2: Steps of the ESG-Consultant pipeline. The user inputs either key company parameters (country, number of employees, etc.) or a natural language question. As the output, the user receives a $\{yes, no, unclear\}$ answer to the applicability question with reasoning and references or a natural-language answer with the exact regulation citation.

2.3 ESG and Artificial Intelligence

There exists a large amount of research in the intersection of Artificial Intelligence (AI) and ESG that analyze and understand the performance of companies related to ESG goals (Sokolov et al., 2021; Pasch and Ehnes, 2022) or use varied machine learning techniques and transformers to assign "ESG scores" or assess the ESG impact of companies using companies' documentation and mining social media data (Kannan and Seki, 2023; Fischbach et al., 2022; Qiu et al., 2023). Furthermore, Nugent et al. (2021); AlNajjar et al. (2024); Wang et al. (2023) do topic classification, modeling, and create taxonomies, to better understand ESG relevant documents provided by companies.

However, to the best of our knowledge, none of the works apply AI/NLP to ESG regulations to solve the tasks we examine in our work.

3 System Design

ESG-Consultant is designed to be used in industry to assist experts in their day-to-day activities, and thus, increase efficiency in understanding and complying with ESG regulations. Figure 1 displays the home page of the system. The user is presented with a short introduction and the list of the system modules. Each of them can be accessed by clicking on their name on the left side panel.

In the following subsections, we discuss each tool of the system in more detail and also describe the pipeline architecture behind the frontend.

Explore regulations module allows the user to select a regulation from a drop-down list of stored regulations. One of the premises of our system is that the system administrator will handle and curate the regulations available for selection. Once a regulation is selected the summary of requirements and penalties for the selected regulation is shown to the user in the form of bullet points, one bullet point per requirement or penalty. In this screen, the user can also access the non-summarized requirements and penalties, which we call "details". In this way, the user can both have an overview of requirements and penalties and access the exact texts from the law for a more in-depth analysis.

Run an evaluation module allows the user to select an ESG regulation, this time to assess whether this regulation applies to a company given the company's key parameters. In this screen from Figure 3, we present to the user a form to input 12 company key parameters, the company name, and give a name to this particular evaluation.

Once the form is filled up the user can hit the "Run evaluation" button. If the form is successfully validated, a method runs asynchronously to perform the applicability task over all key parameters using our ESG Law Applicability pipeline. Since this operation may take some time, the user is given continuous feedback on what key parameters have been evaluated and how many are remaining. Once the evaluation is finished, it is saved to the database, and a link to view the result is shown to the user at the top of the screen. This link leads to the *Saved evaluation* module.

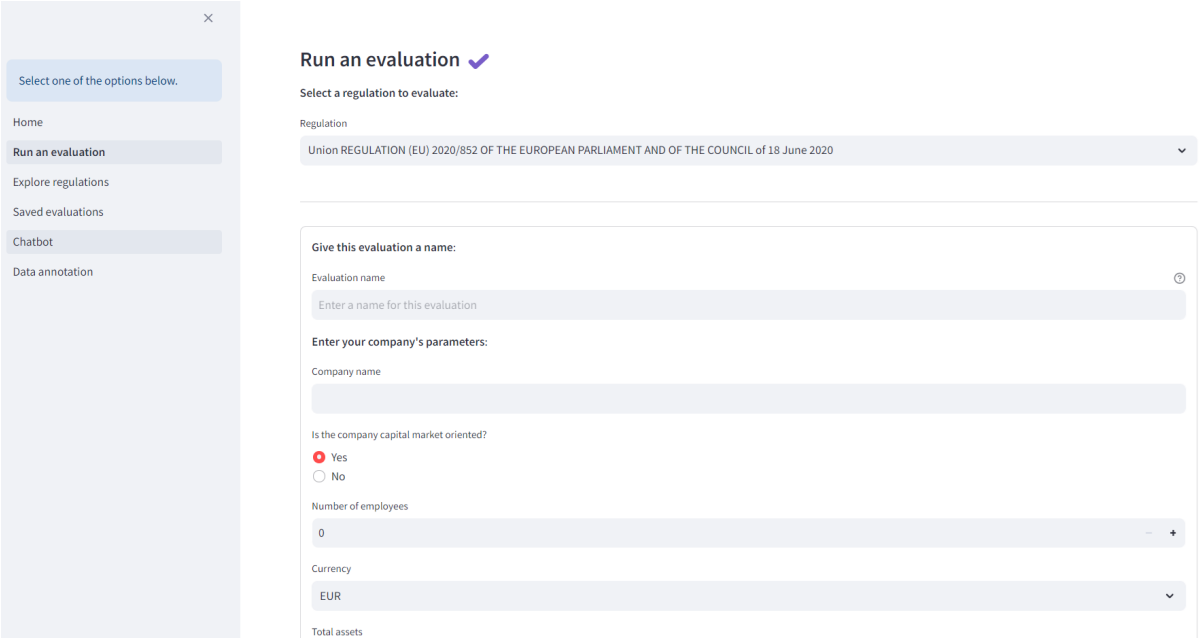

Figure 3: Screenshot of "Run Evaluation" module of ESG-Consultant.

Saved evaluation module shows all previously saved results of the ESG Law Applicability task. The user can select from a table an evaluation result to view. When a result is selected, the details of the evaluation are shown to the user for each key parameter. These details are: The key parameter name and value, the applicability decision $\{yes, no, unclear\}$, reasoning explaining the decision, and three exact excerpts from the law supporting the decision. With this information, the user can not only see the results but verify them using the exact text from the regulations.

Chatbot module (Figure 6 in Appendix A) presents a user-friendly chatbot interface to ask questions about ESG regulations. The user must first select what regulations they want to ask questions about from a drop-down list. Once they type in the question in the text box and hit send, we run our ESG Question Answering pipeline and return exact excerpts as the response from the ESG regulations. We do this following the results of Cheong et al. (2024), whose study showed that legal professionals preferred non-opinion facts as responses.

Data Annotation module aims to carry out a data annotation task to gather a dataset for evaluating the ESG Law Applicability task. This module can continue to be used to gather further annotations for additional experiments. The annotators access this module with a granted password. According to Figure 7 in Appendix A, they are presented with the motivation of the task and are provided with detailed instructions. The task is to answer whether this regulation applies to the company based on the provided key parameter and the relevant excerpts. Since we use data obtained from our ESG applicability pipeline, we also check whether the user's answer coincides with our system's answer (not shown to the user). Then we also present to the user a reasoning for the answer and ask the user to rate this reasoning on a scale from 1 to 5 (with 5 being the best score).

3.1 System Architecture

Our system is developed in the Python programming language, and we use Poetry[2] as our dependency manager. To develop the user interface of our ESG-Consultant for ESG compliance we use streamlit[3]. It is an open-source Python library for the rapid development of user interfaces for data, and AI applications. We adhere to the SOLID (Martin, 2017) principles of software design whenever it aligns with our architecture goals and constraints. The backend can be split into two modules, a retrieval module, and a generation module. Figure 2 demonstrates the whole architecture pipeline.

In the retrieval stage (steps 1 through 4), we have the company's key parameters (1A) (which are added to the specific search prompt) or natural language questions (1B) as user input. Both

[2] https://python-poetry.org/
[3] https://streamlit.io/

Question	Answer
What is the definition of sustainability risk?	'sustainability risk' means an environmental, social or governance event or condition that, if it occurs, could cause an actual or a potential material negative impact on the value of the investment;
Who should participate in the elaboration of waste management plans and waste prevention programmes?	Member States shall ensure that relevant stakeholders and authorities and the general public have the opportunity to participate in the elab oration of the waste management plans and waste prevention programmes.

Table 1: Example question-answer-prediction triplets from our ESG Law Question Answering dataset. The word "elaboration" was separated during pdf text extraction.

are further transformed (step 2) into dense vector representations using an embedding model. In step 3, the search prompt or question text and its dense vector representation are used to perform a hybrid search (keyword and vector) with reciprocal rank fusion[4] for relevant documents in our knowledge base, which contains previously stored ESG documents. We retrieve between 9-18 relevant documents (step 4) from our database. Then we use the cross-encoder model bge-reranker-v2-m3 (Chen et al., 2024) to rerank and keep the top three most relevant documents. We use these techniques because several studies have confirmed task performance gains from hybrid search (Arivazhagan et al., 2023; Gao et al., 2021; Zhang et al., 2023a; Kuzi et al., 2020) and reranking (Boytsov et al., 2022; Nogueira and Cho, 2019) for retrieval. In contrast to our final pipeline naive RAG uses vector search only and no reranking.

In the generation stage (steps 5 through 6), we have either a prompt template for one of the key parameters for the company and its value or a natural language question and the top three relevant text chunks to be used as context.

Finally, we use a single database (PostgreSQL[5]) to store embeddings of 8 ESG regulations. 7 of them are European Union law and were collected from EURLex[6] in text PDF format. The remaining one corresponds to Indian law[7] with a different text layout and consisting of a PDF with scanned images. The regulations' titles with their corresponding number of pages can be seen in Table 5 in Appendix A. We collect the texts using Unstructured[8] and divide them into chunks of preferred size in characters of 1024, max size 2048, and ended with an average character size over all chunks of 1201. The average number of chunks over all regulations is ≈ 114. From Table 5 it can be also seen that the legislation sizes are rather long, therefore, approaches with longer contexts are not possible.

4 ESG Law Question Answering

The ESG Law Question Answering task is a domain-specific task that takes a natural-language question about ESG regulations as an input, expecting a natural-language answer as the output. To solve this task, we use the RAG pipeline which accepts a question as input, retrieves relevant excerpts, and using them in a prompt to the LLM, responds grounded on the retrieved excerpts.

Dataset There is no existing dataset for ESG legislation, therefore, we generate it using GPT-4[9], following a procedure similar to that of de Luis Balaguer et al. (2024). First, we go through all chunks for each regulation, retrieving three chunks at a time as context to generate the questions. For each set of text chunks we ask the model to generate two question-answer pairs with a prompt inspired by the prompt in LangChain QAGenerateChain[10],

[4] https://learn.microsoft.com/en-us/azure/search/hybrid-search-ranking

[5] https://www.postgresql.org

[6] "EUR-Lex" accessed 05.07.2024, https://eur-lex.europa.eu/homepage.html

[7] "Indian Standard Guidelines for Recycling of Plastics" accessed 05.07.2024, https://law.resource.org/pub/in/bis/S11/is.14534.1998.pdf

[8] https://unstructured.io/

[9] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[10] https://api.python.langchain.com/en/latest/evaluation/langchain.evaluation.qa.generate_chain.QAGenerateChain.html

adapted for our task. We specify in the prompt the topics we are interested in, based on system requirements, and ask for a variety of questions using distinct question words[11]. Examples of question-answer pairs are provided in Table 1.

We normalize the texts by removing consecutive spaces and making them lowercase. We remove pairs whose answer is not an exact copy from the text chunks and pairs whose answer is a combination of multiple text chunks, as we want to limit the answer to be contained to a single text chunk for evaluation. Finally, we do a random split of the question-answer pairs in dev and test datasets.

As a result, the development set comprises 273 question-answer pairs (327 is the average answer size in characters); the test set consists of 263 pairs (330 characters on average).

Methodology For this task, we use question inputs directly from a user/dataset. Once a question is asked, we retrieve the top three most relevant text chunks from the database through similarity search techniques. In the generation stage, we have a single prompt template in which the retrieved context and the question are injected. The LLM-generated output is considered to be the final answer to the natural-language input question.

We compare two pipelines: naive (only the basic components of RAG with semantic search) and final (with the hybrid search with Reciprocal Rank Fusion (RRF) and a cross-encoder reranking module). We also experiment with two models: GPT-4 and SaulLM-7b (an LLM trained on legal texts) (Colombo et al., 2024). The number of retrieved documents k varies from $k = 1$ to $k = 3$.

Evaluation Metrics used for the current study are standard: MAP, MRR, and Hit Rate for Retrieval Evaluation (Gao et al., 2023); SQuAD metrics (Exact Match (EM), Precision, Recall, F1) (Rajpurkar et al., 2016) for the whole pipeline.

Results and Discussion Table 2 shows the comparative performance of naive RAG against our final RAG pipeline for GPT-4 and SaulLM-7b on dev set. The final pipeline yields a few percentage points improvements over the naive RAG pipeline. Similarly, we still observe that GPT-4 has the best performance for all metrics, therefore we stick to it for later experiments on the test dataset and ESG Law Applicability Task.

[11] https://dictionary.cambridge.org/grammar/british-grammar/question-words

Model	EM	F1	Precision	Recall
dev set				
GPT4-0613 (n)	**45.42**	73.94	71.55	85.03
GPT4-0613	44.69	**77.29**	**74.08**	**90.23**
SaulLM-7b (n)	10.26	63.34	64.61	73.92
SaulLM-7b	13.92	67.63	67.79	78.97
test set				
GPT4-0613	**44.86**	**78.87**	75.51	**92.02**
SaulLM-7b	14.83	72.90	**76.32**	79.13

Table 2: SQuAD metrics of GPT-4 and SaulLM-7b models on the ESG Law Question Answering on dev and test datasets. The results are shown for the naive RAG approach *(n)* and the final pipeline.

@k	MAP	MRR	Hit Rate
1	0.8479	0.8479	0.8479
2	0.8878	0.8878	0.9278
3	0.8945	0.8942	0.9468

Table 3: Retrieval metrics on the ESG Law Question Answering on the test dataset with varying the number of retrieved documents k from $k = 1$ to $k = 3$.

Another interesting finding is that the SQuAD recall for GPT4 in our final pipeline achieves a value of 90%, this means the correct answer is within the model answer in 9 out of 10 questions. Figure 4 in Appendix A demonstrates that the SQuAD recall value with our final pipeline is close to the calculated upper bound. We interpret this as the models being reasonably good at finding the answers within the context, while the answers provided tend to be longer than the ground truth. The rest of the metrics still show a wide gap to the upper bound.

Finally, we evaluate our final pipeline on the test split of our ESG Law Question Answering dataset. The results for the SQuAD metrics can be seen in Table 2, and the results for the retrieval metrics in Table 3. The results on the previously unseen test dataset are coherent with the development dataset.

Hallucination Evaluation To evaluate hallucinations in large language models (LLMs), we employed an open source hallucination evaluation model called Lynx (Ravi et al., 2024), a state-of-the-art system designed for advanced reasoning in real-world scenarios. Lynx detects hallucinations by comparing the predicted answer generated by the LLM with the context provided with the input question. If the predicted answer is not supported by the context when considered alongside the question, the model is deemed to be hallucinating.

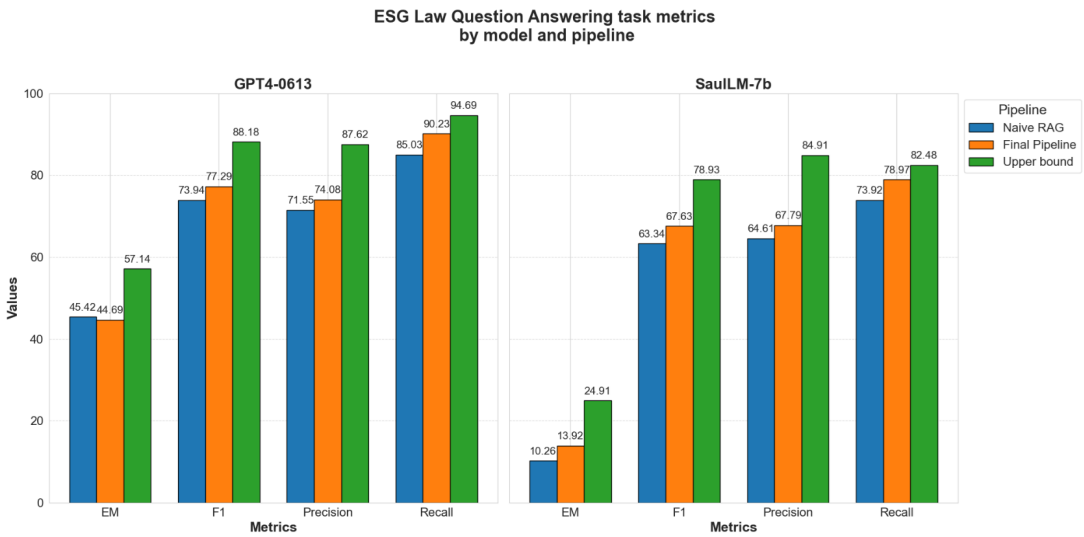

Figure 4: ESG Law Question Answering metrics comparing the results of naive RAG, final pipeline, and upper bound performance for both GPT4-0613 and SaulLM7-b. The upper bound was calculated assuming a perfect retriever, we pass only the text chunk containing the ground truth answer to the model as context.

Using Lynx, we evaluated hallucination rates for both LLMs predictions: GPT4-0613 and SauLM-7b. The results demonstrate a low incidence of hallucinations: 260 out of 263 passes (98.86% accuracy) for GPT-4 and 251 out of 263 passes (95.43% accuracy) for SauLM-7b. These high pass rates suggest that both models perform well in avoiding hallucinations under this evaluation framework, with only a few failures identified in each case.

To further understand the nature of low scores, a detailed manual analysis is conducted for the best-scored GPT-4 model. First, we identify the number of cases where predicted answer does not have the exact match (142 out of 263 instances). Second, we count how many times the true answer cannot be found in the predicted answer (67 out of 263 instances). Third, in 63 out of 263 instances, predicted answer cannot be found in the provided context. 36 cases is the intersection between the second and the third case.

The manual review reveals that many of these discrepancies are not the hallucinations but rather inaccurate text reproduction (see 6). In several cases, LLMs omit parts of the answer that it considers to be irrelevant to the question, leading to a mismatch. Additionally, artifacts such as separated words ("sep arated") cause string mismatches, even though the semantic meaning of the prediction remained consistent with the context.

These findings highlight that while Lynx effectively identifies hallucinations, some of the identified cases were due to formatting or textual omissions. Future improvements could focus on either refining the evaluation criteria to account for these minor deviations or implementing better solutions to reproduce the exact matches.

5 ESG Law Applicability Task

The ESG Law Applicability task aims at determining whether a specific regulation applies to a company based on its parameters or characteristics. This is a classification task with three possible classes: "yes", "no", and "unclear", evaluated with the traditional F1 score.

To build and evaluate a system to perform the law applicability task, twelve key parameters were identified to be used as criteria for the applicability of the ESG regulations. The list of the key parameters identified by legal experts is present in Table 4.

Methodology We hypothesize that by providing the LLM access to the law text and the key parameter values for a company via the RAG pipeline, it should be able to determine if a particular ESG regulation applies to the company. Therefore, for each parameter, we use static pre-made search prompts to retrieve relevant excerpts. Then, the GPT-4 model is fed with the parameter keys and values as well as the retrieved texts to generate the answer.

Dataset Annotation The ESG Law Applicability dataset was created through an annotation task by three European experts familiar with ESG law (annotators). Using the *Data Annotation* module in our ESG-Consultant, the annotators were shown

1. Is a capital market-oriented company (listed on the stock exchange)	7. Is offering financial products
2. Number of employees	8. Jurisdiction (countries)
3. Assets	9. Markets (countries)
4. Yearly revenue	10. Sourcing (countries)
5. Is in the Scope of the Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH)	11. Production (countries)
6. Is a manufacturer or distributor of batteries	12. Products and services offered

Table 4: Twelve key parameters identified to be used as criteria for the applicability of the ESG regulations.

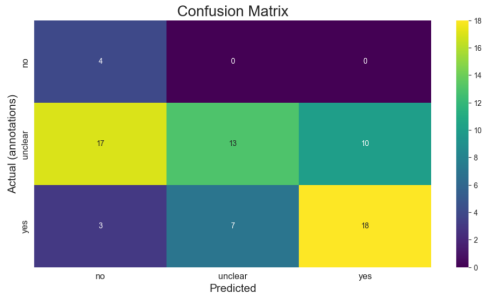

Figure 5: Confusion matrix of targets against system predicted values. The main diagonal represents correct classifications. Most of the mismatches occur for the "$unclear$" label.

36 law applicability task results. Annotators were asked to judge if the ESG regulation applies to the company based on the key parameter value and the relevant law texts (see Appendix C for more detail). If their answer coincided with the one from our system, then they were shown the reasoning generated by the system and asked to rate it on a scale of 1 to 5 (with 5 being the best score). They were also asked whether each of the law texts shown was relevant to answering the law applicability question.

The next step is to assess the inter-rater reliability of the applicability task annotations using Cohen's kappa (Cohen, 1960) and Krippendorff's alpha (Hayes and Krippendorff, 2007), which was only 0.15 when considering three labels and grew up to 0.81 when considering $\{yes, no\}$ labels only.

Results and Discussion To evaluate the ESG Law Applicability task we use the dataset collected during the above-mentioned annotation step. We use the results of the annotations as the ground truth dataset to calculate standard classification metrics against the values predicted by our system.

The average macro-F1 results are 0.65 for three classes $\{yes, no, unclear\}$, which is rather low. Therefore, we further examined in what instances the target and predicted values differ. For this, we created a confusion matrix of the target against the predicted values, see Figure 5.

We see that the majority of mismatches between the annotated and predicted values occur for the $unclear$ label, stating that these cases may be especially hard. For the cases with a definitive answer, we see a higher agreement between the model and annotators. Therefore, we exclude the tasks that got an $unclear$ labeling by either the model or the annotators. The recalculated metrics for labels $\{yes, no\}$ achieve F1 of 0.92 (the number of samples reduced from 72 (36 pairs) to 25).

We observed a much higher performance of our system when considering only definitive $\{yes, no\}$ answers for the ESG Law Applicability task. Even though our dataset is too tiny to make generalizations, we still can observe that this is consistent with our findings when evaluating the inter-rater reliability for the annotation task: the agreement was high when we excluded tasks with the $unclear$ labels. This leads us to question the reliability of the $unclear$ annotations in our dataset. A new annotation task and further experiments are required to resolve this issue.

User Study In addition to the labelling task, we ask annotators to evaluate the reasoning of the system answers, when the answer of the system and annotator matched. We ask them to rate our system's provided reasoning for the answer on a scale from 1 to 5 (with 5 being the best score). The average rating is 4.28, meaning that the annotators find the reasoning a good explanation for the decision.

6 Conclusion

In this paper, we developed the ESG-Consultant system for ESG compliance based on RAG pipeline and specified two tasks that these experts considered highly valuable for assisting in ESG compliance: the ESG Law Question answering task, and the ESG Law Applicability task. Moreover, we created two novel datasets and evaluated our baselines for both tasks. In future work, we plan to collect more data and work on multi-step reasoning.

Limitations

The main limitations of the paper are as follows:

- First, the space of implementations for RAG is huge and we could test only a small subset we considered the most appropriate for this research. The current study lacks more additional baselines to comparre the model with. The main contributions of the paper are the demonstration system and the datasets, therefore, we have decided to keep the number of tested approaches limited. We plan to improve this in future work.

- ESG Law Question Answering dataset is generated automatically, which could have lead to unrelated and not reflective data of the types of questions an expert would formulate. To approach this limitation, one legal expert was asked to verify some random question-answer pairs in order to approve the relevance of the dataset. However more analysis should be done as future work.

- The size of the ESG Law Applicability Task is very small due to expensive and time-consuming annotation process. Involving legal experts in the training and evaluation process can help to refine the model, ensuring that it properly reflects the understanding of the task by legal experts. Unfortunately, we could only have access to a group of three annotators familiar with ESG regulations. Furthermore, we could only use a few hours of their time, which led to us running the annotation project in one pass with 36 tasks that were distributed in pairs to annotators, for a total of 72 samples. Relying on such a small number of samples might not reflect the overall performance of the system and be not as representative as it could be with a larger number of examples. Nevertheless, legal questions are always based on interpretation, which might differ from expert to expert and result in a low inter-annotation agreement, which we have already seen during our research.

- The accuracy and reliability of the model are critical; moderate results might indicate that the model may not capture the nuances and complexities of ESG legal concepts.

- One may doubt the methodological soundness to employ GPT-4 for both generating data and subsequently assessing that data. However, nowadays, many approaches employ GPT-4 for these tasks subsequently, e.g. the study of Guan et al. (2024) on fact verification.

The shortcomings should be taken into consideration when deploying our system to real-world applications. Nevertheless, we believe that this work still can help practitioners build assisting technology for legal professionals. We believe that this is an important application field, where more research should be conducted to improve legal services.

Ethical Statement

When developing a system to assist law experts with ESG regulations, several ethical considerations must be addressed.

- Even though we evaluate the bias which could be produced within the system, one should still care about bias and hallucination of the model, as it could be extremely harmful to the legal domain. Therefore, we do not recommend using such a system without human supervision.

- In our research, we used closed-source models. On the one hand, they charge per token which makes running the experiments prohibitively expensive. On the other hand, when using closed models, we have limited control over the model and its training data, making customization and fine-tuning more difficult.

- One must consider the potential misuse of the model. If the model is not highly accurate, relying on its predictions without human verification could lead to misinterpretations of legal elements, which in turn could have policy implications or affect legal interpretations and decisions.

- Our ESG regulations are collected for European countries and Inida, therefore, the users should take care about the application of these legislations for other countries.

References

Sallam Abualhaija, Chetan Arora, and Lionel C. Briand. 2022. COREQQA - A compliance requirements understanding using question answering tool. *CoRR*, abs/2206.10233.

Saif Majdi AlNajjar, Xinyu Wang, and Yulan He. 2024. Topic taxonomy construction from ESG reports. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

Manoj Ghuhan Arivazhagan, Lan Liu, Peng Qi, Xinchi Chen, William Yang Wang, and Zhiheng Huang. 2023. Hybrid hierarchical retrieval for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10680–10689.

Leonid Boytsov, Tianyi Lin, Fangwei Gao, Yutian Zhao, Jeffrey Huang, and Eric Nyberg. 2022. Understanding performance of long-document ranking models through comprehensive evaluation and leaderboarding. *arXiv preprint arXiv:2207.01262*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. 2024. (a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2454–2469, New York, NY, USA. Association for Computing Machinery.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, André F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. SaulLM-7B: A pioneering large language model for law. *CoRR*, abs/2403.03883.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

FeiFei Dai, ZhangLi Zhao, ChangPeng Sun, and Bo Li. 2022. Intelligent audit question answering system based on knowledge graph and semantic similarity. In *2022 11th International Conference of Information and Communication Technology (ICTech))*, pages 125–132.

Maria Angels de Luis Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. 2024. RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *CoRR*, abs/2401.08406.

ESG. 2023. ESG Policy Digest: December 2023. `https://www.esgbook.com/esg-policy-digest-december-2023/`.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. LawBench: Benchmarking legal knowledge of large language models. *CoRR*, abs/2309.16289.

Jannik Fischbach, Max Adam, Victor Dzhagatspanyan, Daniel Méndez, Julian Frattini, Oleksandr Kosenkov, and Parisa Elahidoost. 2022. Automatic esg assessment of companies by mining and evaluating media coverage data: Nlp approach and tool. *2023 IEEE International Conference on Big Data (BigData)*, pages 2823–2830.

Debasis Ganguly, Jack G. Conrad, Kripabandhu Ghosh, Saptarshi Ghosh, Pawan Goyal, Paheli Bhattacharya, Shubham Kumar Nigam, and Shounak Paul. 2023.

Legal IR and NLP: the history, challenges, and state-of-the-art. In *European Conference on Information Retrieval*, volume 13982 of *Lecture Notes in Computer Science*, pages 331–340. Springer.

Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement lexical retrieval model with semantic residual embeddings. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*, pages 146–160. Springer.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Stuart Gillan, Jay C Hartzell, Andrew Koch, and Laura T Starks. 2010. Firms' environmental, social and governance (esg) choices, performance and managerial motivation. *Unpublished working paper*, 10.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111, Mexico City, Mexico. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Christoph Hoppe, David Pelkmann, Nico Migenda, Daniel Hötte, 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.

Ngo Quang Huy, Nguyen Manh Duc Tuan, Nguyen Anh Duong, and Pham Quang Nhat Minh. 2021. Aimelaw at alqac 2021: Enriching neural network models with legal-domain knowledge. In *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for ESG scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 45–54, Macao.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Saar Kuzi, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach. *ArXiv*, abs/2010.01195.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Ting-Ting Li, Kai Wang, Toshiyuki Sueyoshi, and Derek D Wang. 2021. ESG: Research progress and future prospects. *Sustainability*, 13(21):11663.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.

R Martin. 2017. *Clean Architecture: A Craftsman's Guide to Software Structure and Design*, 1st edition. United States of America: Pearson Education Inc, USA.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. 2024. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, 382(2270):20230159.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Tim Nugent, Nicole Stelea, and Jochen L Leidner. 2021. Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation. In *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings 14*, pages 157–169. Springer.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Stefan Pasch and Daniel Ehnes. 2022. Nlp for responsible finance: Fine-tuning transformer-based models for esg. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3532–3536.

Rajvardhan Patil and Venkat Gudivada. 2024. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5).

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

Le Qiu, Bo Peng, Jinghang Gu, Yu-Yin Hsu, and Emmanuele Chersoni. 2023. Identifying ESG impact with key information. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 51–56, Bali, Indonesia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. Lynx: An open source hallucination evaluation model. *CoRR*, abs/2407.08488.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore. Association for Computational Linguistics.

Alik Sokolov, Jonathan Mostovoy, Jack Ding, and Luis A. Seco. 2021. Building machine learning systems for automated esg scoring. In *The Journal of Impact and ESG Investing*.

Anu Thomas and S Sangeetha. 2022. Knowledge graph based question-answering system for effective case law analysis. In *Evolution in Computational Intelligence: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, pages 291–300. Springer.

Xinyu Wang, Lin Gui, and Yulan He. 2023. A scalable framework for table of contents extraction from complex ESG annual reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13215–13229, Singapore. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Peitian Zhang, Zheng Liu, Shitao Xiao, Zhicheng Dou, and Jing Yao. 2023a. Hybrid inverted index is a robust accelerator for dense retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1877–1888, Singapore. Association for Computational Linguistics.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023b. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.

A System Screenshots

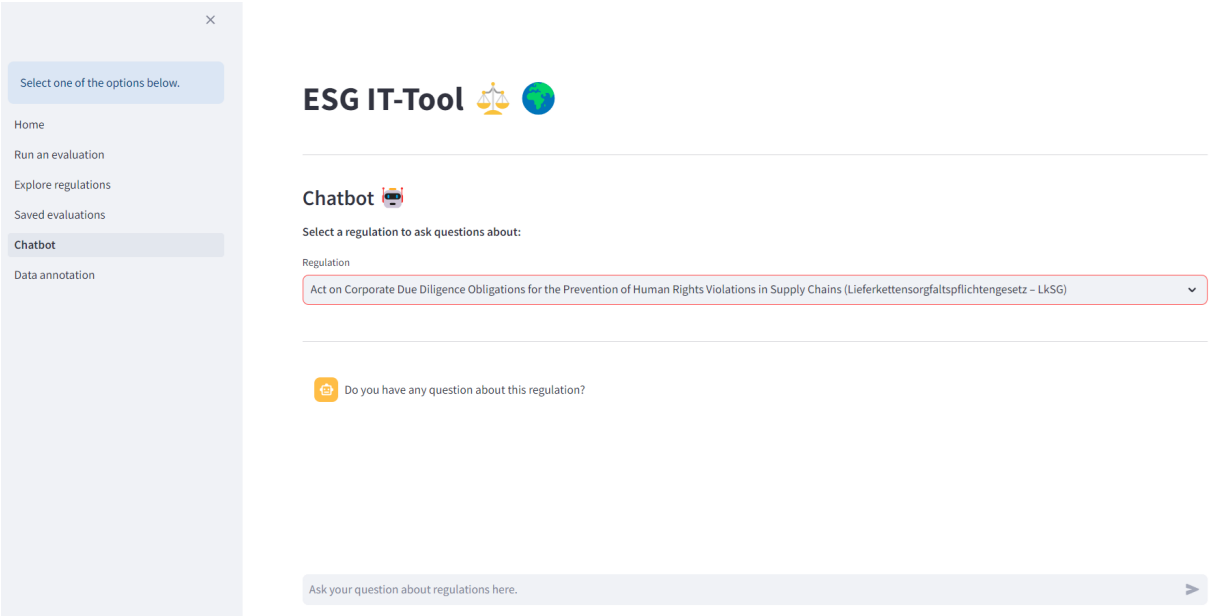

Figure 6: Screenshot of the Chatbot Module of the ESG-Consultant.

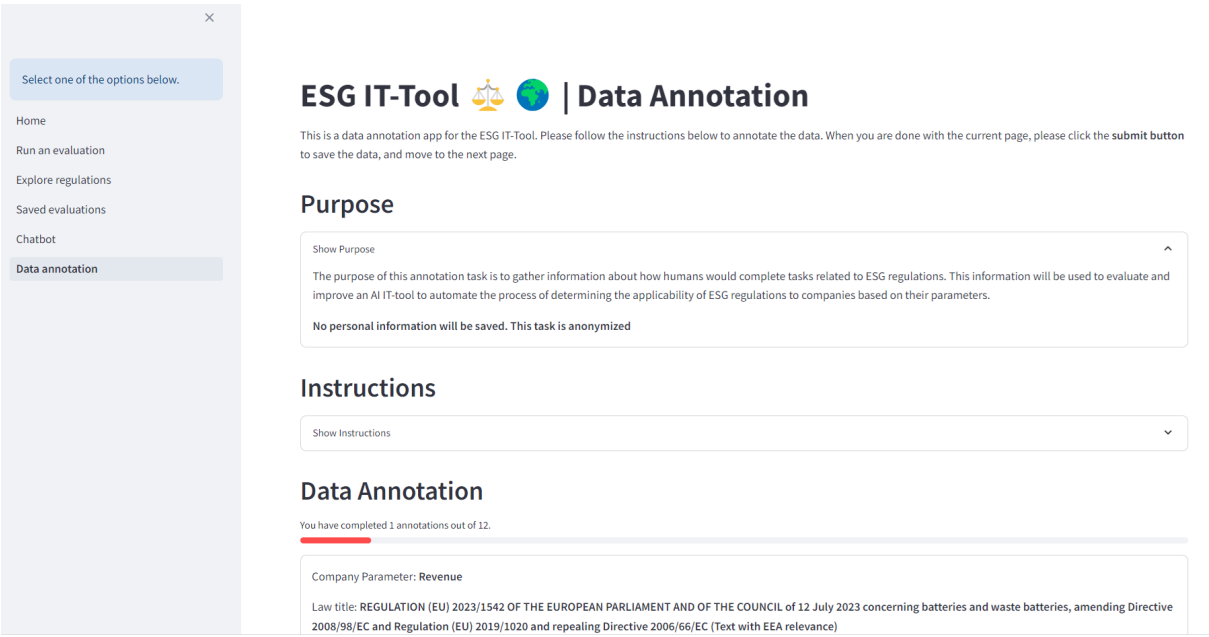

Figure 7: Screenshot of the“Annotation” Module of ESG-Consultant.

B ESG Regulations List

Title	Page Count
REGULATION (EU) 2023/1542 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 12 July 2023 concerning batteries and waste batteries, amending Directive 2008/98/EC and Regulation (EU) 2019/1020 and repealing Directive 2006/66/EC (Text with EEA relevance)	117
DIRECTIVES DIRECTIVE (EU) 2022/2464 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL	66
DIRECTIVE 2008/98/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 19 November 2008 on waste and repealing certain Directives (Text with EEA relevance)	59
Union REGULATION (EU) 2020/852 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 June 2020	31
EXPLANATORY MEMORANDUM COMMISSION DELEGATED REGULATION (EU) of 6.7.2021 supplementing Regulation (EU) 2020/852 of the European Parliament and of the Council	28
Act on Corporate Due Diligence Obligations for the Prevention of Human Rights Violations in Supply Chains (Lieferkettensorgfaltspflichtengesetz – LkSG)	22
REGULATION (EU) 2019/2088 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 November 2019 on sustainability-related disclosures in the financial services sector (Text with EEA relevance)	16
Indian Standard GUIDELINES FOR RECYCLING OF PLASTICS	9

Table 5: Titles of ESG regulations used for this work with their respective page count.

C Annotation Guidelines

```
The purpose of this annotation task is to gather information about how humans would complete
tasks related to ESG regulations. This information will be used to evaluate and improve
an AI IT-tool to automate the process of determining the applicability of ESG regulations
to companies based on their parameters. No personal information will be saved.
This task is anonymized.

You will be presented with company parameters and law text extracted from ESG regulations.
In each page you will be asked to complete some tasks related to a company parameter
(i.e number of employees, revenue, etc). When you complete the tasks in the current page,
please press submit or submit rating at the bottom of the page. Once each page is submitted
it cannot be changed. Answer to the best of your capacity but don't overthink too much
or spend a lot of time on a page.

Here are the tasks you will be asked to complete in each page:

 1. Read carefully the text extracted from ESG law, and decide if the text is relevant for
the current parameter. The text is relevant if we could use it to determine if this law
(see law title) applies to a company given the current parameter.

 2. Considering all three law texts you read in the current page, and the company parameter
value, decide if this law applies to the company based on the parameter value.

 3. In some pages, you will be presented a reasoning explaining why the law applies to a
company or not for the current parameter value. Your task is to rate this reasoning on
a scale of 1 to 5, where 1 is the lowest and 5 is the highest.
```

Figure 8: Annotation guidelines for the ESG Law Applicability dataset collection.

Figure 9 shows an example of asking annotators to assess the applicability of a regulation. Since their answer coincided with the decision of our system (not shown to the annotator). We ask the annotator to rate the system's reasoning.

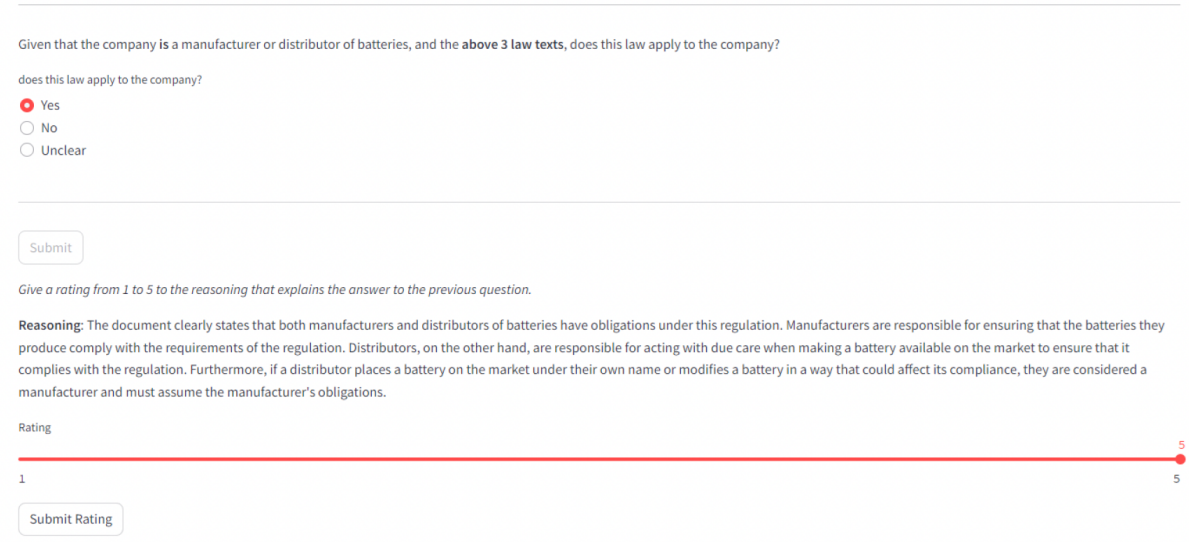

Figure 9: Example of asking annotators to assess the applicability of a regulation. Since their answer coincided with the decision of our system (not shown to the annotator). We ask the annotator to rate the system's reasoning.

D Experimental Setup and Hyperparameters

The experiments utilized Nvidia RTX A3000 with 27.8GB of GPU memory during inference. In total during experimentation we used approximately 120h of GPU time. A budget of 350$ was spent on the GPT-4 computations and 540$ on Azure infrastructure over the course of the research project.

For our experiments we didn't perform finetuning of models and instead explored distinct RAG pipelines. On the retrieval stage we used OpenAI's text-ada-002 embeddings[12] for vector search, for hybrid search we added PostgreSQL's Full-text search[13]. To fuse hybrid search results we explored Reciprocal Rank Fusion (RRF)(Cormack et al., 2009)

$$RRFscore(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)}$$

Additionally, we explored Relative Score Fusion (RSF)[14]. Where each score $s \in S$, obtained from the search methods, is mapped to a normalized score $s_n \in [0, 1]$ by using a MinMax scaler.

$$s_n = \frac{s - s_{min}}{s_{max} - s_{min} + \epsilon}$$

We add a small constant ϵ to the denominator to avoid division by zero. These normalized scores are then summed up, and divided by the total number of available scores.

$$RSFscore(d \in D) = \sum_{s_n \in S_n} \frac{s_n}{|S_n|}$$

We also added a cross-encoder model for reranking, bge-reranker-v2-m3 (Chen et al., 2024).

On the generation stage we used SaulLM-7b (Colombo et al., 2024) and GPT4-0613 accessed through a deployment in Microsoft Azure [15].

Our best performing pipeline consisted of hybrid search with Reciprocal Rank Fusion, cross-encoder reranking module, and GPT4-0613.

[12] https://openai.com/index/new-and-improved-embedding-model/
[13] https://www.postgresql.org/docs/current/textsearch-intro.html
[14] https://weaviate.io/blog/hybrid-search-fusion-algorithms
[15] https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models?tabs=python-secure#gpt-4

E Additional examples

Question	Answer	Prediction
What is the purpose of Directive 2008/105/EC?	Directive 2008/105/EC of the European Parliament and of the Council of 16 December 2008 on environmental quality standards in the field of water policy, amending and subsequently repealing Council Directives 82/176/EEC, 83/513/EEC, 84/156/EEC, 84/491/EEC, 86/280/EEC and amending Directive 2000/60/EC of the European Parliament and of the Council (OJ L 348, 24.12.2008, p. 84) .	Directive 2008/105/EC of the European Parliament and of the Council of 16 December 2008 on environmental quality standards in the field of water policy, amending and subsequently repealing Council Directives 82/176/EEC, 83/513/EEC, 84/156/EEC, 84/491/EEC, 86/280/EEC and amending Directive 2000/60/EC of the European Parliament and of the Council.
When is the application of the delegated act referred to in paragraph 2 expected?	ensuring its application from 1 January 2023.	ensuring its application from 1 January 2023. The Commission shall adopt the delegated act referred to in paragraph 2 by 31 December 2021, with a view to
What does Regulation (EU) No 995/2010 pertain to?	Regulation (EU) No 995/2010 of the European Parliament and of the Council of 20 October 2010 laying down the obligations of operators who place timber and timber products on the market	Regulation (EU) No 995/2010 of the European Parliament and of the Council of 20 October 2010 lays down the obligations of operators who place timber and timber products on the market .

Table 6: Example question-answer-prediction triplets from our ESG Law Question Answering dataset. The word "elaboration" was separated during pdf text extraction. Missed or extra text is marked in yellow .