





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

Angel Ontiveros^{1,2}, Irina Nikishina¹ , Moritz Gomm²,
Christopher Schmitt², and Chris Biemann¹ 

University of Hamburg, Hamburg, Germany
{angel.ontiveros irina.nikishina chris.Biemann}@uni-hamburg.de
² Zühlke Engineering GmbH, Eschborn, Germany
{moritz.gomm christopher.schmitt}@uni-hamburg.de

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 discuss 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 (https://github.com/angelonti/ai4esg_experiments, <https://ai4esg-app.azurewebsites.net/> (pass1: e2c306nt for the annotations), <https://www.youtube.com/watch?v=JKKGD3lFiUA>).

Keywords: ESG Compliance · Retrieval Augmented Generation · Legal Data · Demonstration System

Introduction

This paper focuses on the Environmental, Social, and Governance (ESG) regulations which have become pivotal benchmarks in contemporary corporate assessment [8]. 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 [3]. 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 Large Language Models (LLMs) 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 provide support for two tasks in this demonstration system: the **ESG Law Question Answering task (ESG-QA)**, a vanilla Question Answering task with the texts from the ESG domain and the **ESG Law Applicability task**, a standard classification task, which aims at determining whether a ESG regulation applies to a company with specific parameters.



Fig 1 Home screen of the ESG-Consultant tool.

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 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 present an additional annotation module, for creating manual annotations for ESG regulations, which can be used for further improvement of the system.
- We introduce a novel domain-related benchmark: ESG Law Question Answering, and create a dataset for the evaluation.

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

Due to specifically complex language [7], specialized subdomains [10], and heterogeneous formats [4], the legal domain remains one of the most challenging NLP domains. However, more and more research on legal tasks now makes use of LLMs [4, 7]. This has led to applications of NLP for the law being increasingly adopted in legal settings [6]. However, none of the previous work applied RAG to ESG-QA.

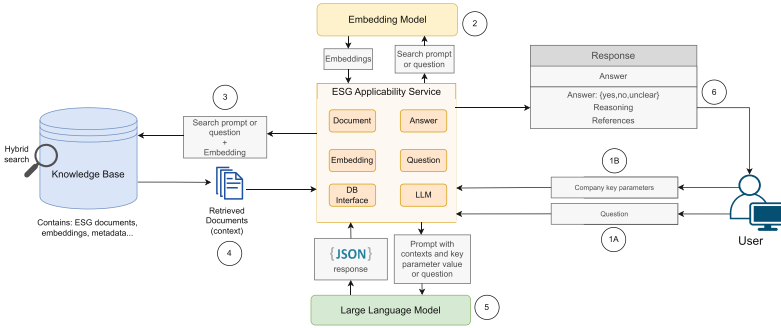


Fig 2 Steps of the ESG-Consultant pipeline. The user inputs either key company parameters (country, number of employees, etc.) or a natural language question. As the output, the user receives a *yes no unclear* answer to the applicability question with reasoning and references or a natural-language answer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

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.

System Architecture. Our system is developed in the Python programming language, and we use Poetry¹ as our dependency manager. To develop the user interface of our ESG-Consultant for ESG compliance we use streamlit². It is an open-source Python library for the rapid development of user interfaces for data, and Artificial Intelligence (AI) applications. We adhere to the SOLID [9] 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.

¹ <https://python-poetry.org/>.

² <https://streamlit.io/>.

We use a single database (PostgreSQL) to store embeddings of eight ESG regulations. Seven of them are European Union law and were collected from EURLex³ in text PDF format. The remaining one corresponds to Indian law⁴ with a different text layout and consisting of a PDF with scanned images. We collect the texts using Unstructured⁵ and divide them into chunks of preferred size in characters of 1 24, max size 2 48, and ended with an average character size over all chunks of 12 1. The average number of chunks is 114.

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 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 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-m3⁶ 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 [5] and reranking [1] 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.

4 Experiments

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. The ESG Law Applicability task aims at determining whether a specific regulation applies to a company based on its parameters or characteristics. To solve both tasks, we use the Retrieval Augmented Generation (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-4⁷. 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). An example of such data is as follows: *What is the*

³ “EUR-Lex” accessed 05.07.2024, <https://eur-lex.europa.eu/homepage.html>.

⁴ “Indian Standard Guidelines for Recycling of Plastics” accessed 05.07.2024, <https://law.resource.org/pub/in/bis/S11/is.14534.1998.pdf>.

⁵ <https://unstructured.io/>.

⁶ <https://huggingface.co/BAAI/bge-reranker-v2-m3>.

⁷ <https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4>.

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.

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) [2]. The number of retrieved documents varies from $n = 1$ to $n = 3$.

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

Table 1 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.

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

Results and Discussion. Table 1 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. 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. Finally, we evaluate our final pipeline on

Table 2 Retrieval metrics on the ESG Law Question Answering on the test dataset with varying the number of retrieved documents k from $k = 1$ to $k = 3$.

@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

the test split of our ESG Law Question Answering dataset. The results for the SQuAD metrics can be seen in Table 1, and the results for the retrieval metrics in Table 2. The results on the previously unseen test dataset are coherent with the development dataset.

5 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 a novel ESG QA dataset and evaluated on it several baselines. In future work, we plan to collect more data and work on multi-step reasoning.

References

1. Boytsov, L., Lin, T., Gao, F., Zhao, Y., Huang, J., Nyberg, E.: Understanding performance of long-document ranking models through comprehensive evaluation and leaderboarding. arXiv preprint [arXiv:2207.01262](https://arxiv.org/abs/2207.01262) (2022)
2. Colombo, P., et al.: SaulLM-7B: a pioneering large language model for law. CoRR [abs/2403.03883](https://arxiv.org/abs/2403.03883) (2024)
3. ESG: ESG Policy Digest: December 2023 (2023). <https://www.esgbook.com/esg-policy-digest-december-2023/>
4. Ganguly, D., et al.: Legal IR and NLP: the history, challenges, and state-of-the-art. In: Kamps, J., et al. (eds.) Advances in Information Retrieval. ECIR 2023. LNCS, vol. 13982, pp. 331–340. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28241-6_34
5. Gao, L., Dai, Z., Chen, T., Fan, Z., Van Durme, B., Callan, J.: Complement lexical retrieval model with semantic residual embeddings. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12656, pp. 146–160. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72113-8_10
6. Guha, N., et al.: LegalBench: a collaboratively built benchmark for measuring legal reasoning in large language models. Adv. Neural Inf. Process. Syst. **36**, 44123–44279 (2024)
7. Katz, D.M., Hartung, D., Gerlach, L., Jana, A., Bommarito II, M.J.: Natural language processing in the legal domain. arXiv preprint [arXiv:2302.12039](https://arxiv.org/abs/2302.12039) (2023)

8. Li, T.T., Wang, K., Sueyoshi, T., Wang, D.D.: ESG: research progress and future prospects. *Sustainability* **13**(21), 11663 (2021). <https://doi.org/10.3390/su132111663>
9. Martin, R.: *Clean Architecture: A Craftsman's Guide to Software Structure and Design*. United States of America: Pearson Education Inc., USA, 1st edn. (2017). <https://dl.acm.org/doi/10.5555/3175742>
10. Shui, R., Cao, Y., Wang, X., Chua, T.S.: A comprehensive evaluation of large language models on legal judgment prediction. In: Bouamor, H., Pino, J., Bali, K. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7337–7348. Association for Computational Linguistics, Singapore (December 2023)