Self-Explanatory Retrieval-Augmented Generation for SDG Evidence Identification

Darío Garigliotti[1][0000−0002−0331−000X]

University of Bergen, Norway - dario.garigliotti@uib.no

Abstract. With the establishment of the Sustainable Development Goals (SDG) framework, practitioners in environmental impact assessment have an increasing requirement to detect relevant information centered on this frame of reference. The task of automatically identifying evidence that supports the project actually addressing a particular SDG target becomes crucial for enabling assessment digitalization across long, heterogeneous documents. In this work, we tackle SDG evidence identification via the well-suited Retrieval-augmented Generation (RAG) approach powered by Large Language Models (LLM). The identified evidence may also support further related tasks in conceptual modeling where reports or parts of their content are to be assigned to entries in a structured resource such as a domain-specific ontology. Beyond the measurement of performance of a series of method configurations on this task, we also assess RAG abilities for making this kind of decisions when the LLM is requested to explain its own mechanisms alongside the answer it generates. Our evaluation resources are made publicly available.

Keywords: Sustainable Development Goals · Retrieval-augmented Generation · Large Language Models · Explainable AI.

1 Introduction

The requirement for addressing Sustainable Development Goals (SDG) [1] has become more prominent across many spheres of human activity. SDGs, established by United Nations to serve as a common ground for high-level policy arcs, have indeed become a frame of reference in a multiplicity of development regulations at all scales and in almost every domain (Del Campo et al., 2020). Distinguishedly, professionals in areas of environmental assessment, ranging from expert assessors to authorities, as well as key personnel on the side of project developers, are regularly faced with the need to process documentation pertaining SDGs. These kind of documents, *environmental (impact) assessment* (EIA) reports, are typically long, and are presented in heterogeneous formats, through which these experts endure the difficult task of finding the relevant information that is needed, in particular, that which addresses a given SDG of interest. Plus, the information need of environmental assessors for detecting textual passages addressing an SDG, and mostly SDG targets as focused, actionable SDG subgoals,

---
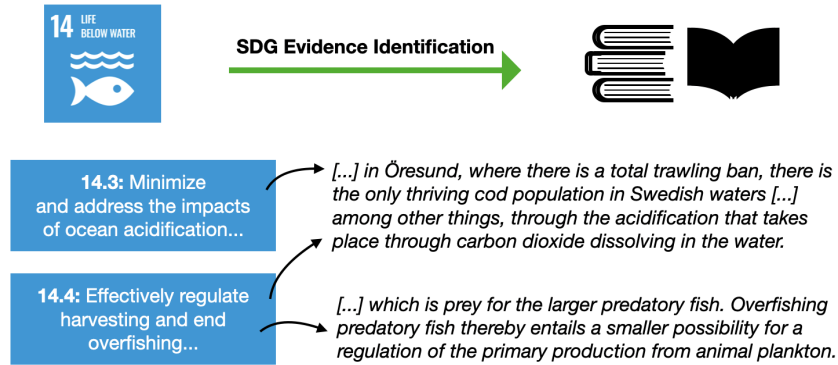
[1] https://sdgs.un.org/goals

Fig. 1: Overview of the SDG Evidence Identification task.

often clashes subtly with incentives in the documents' narrative for aggrandazing the ways by which the SDG target is being addressed by the developer behind the report (Del Campo et al., 2020). There is clearly an increasing need for information systems that enable larger digital access to this kind of analysis by all actors involved. We argue that a crucial component of such a system deals with detecting relevant passages in a EIA report addressing a given SDG target. This information extraction problem is the one that we address in our paper. An overview of this task is depicted in Fig. 1.

A series of recent advancements in key components of their training (architecture, regime, data acquisition and usage) has made significant improvements in language modeling technology with the establishment of Large Language Models (LLMs) as the dominant technology in several information processing tasks (Touvron and et al., 2023; Radford et al., 2019). These models, trained over very large corpora of scrapped content and prominent datasets with powerful transformers-based neural networks and human feedback, provide themselves vast knowledge implicitly stored in their billions of parameters (Elazar and et al., 2024). Yet, for knowledge domains, problems and content sources with limited or altogether null presence in the underlying data of an LLM, the incorporation of explicit, relevant knowledge often results decisive to achieve desired performances. Retrieval-augmented Generation (RAG) (Lewis and et al., 2020) is a general umbrella of methods that integrate external knowledge to complement the capabilities of an LLM. This approach suits very well with our problem, where we intend to predict a decision making answer regarding whether a candidate passage truly addresses an SDG target.

Our work subscribes within the emerging research between areas such as Natural Language Processing (NLP), Knowledge Representation (KR) or Information Extraction (IE) overlapping in the applicability of their techniques on the relevant phenomena within climate change. Whether to respond questionnaires (Spokoyny et al., 2023) or to detect climate claims in text (Stammbach and et al., 2023), more studies address environmental scenarios of information

access. Fundamental research is also focused on building dedicated language models (Webersinke et al., 2022; Thulke and et al., 2024), as well as chat systems (Vaghefi and et al., 2023) and EIA-centric ontologies (Nielsen and et al., 2023; Garigliotti and et al., 2023).

Alongside the convenience of using LLMs in multiple tasks and domains, a common drawback is the presence of hallucinations that very often hurt their performance, and altogether diminish the trustworthiness in decisions supported by these models (Liu et al., 2023; Asai et al., 2022). Hence, in this work, we study mechanisms within Explainable AI (XAI) (Cambria et al., 2023) to elicit, from an LLM, not only the prediction for the correct assessment of a target being addressed by a passage, but also the ability of the same model to explain the rationales behind its generated answer. The potential explanations, together with the evidence naturally provided by the identification task, would increase the factors of trust in the users of these environmental assessment systems (Menick and et al., 2022).

The resources developed in this work are made publicly available at `https://bit.ly/AIMM_at_ER_2024-SDG_EI_XAI-Materials`.

The rest of the paper defines in detail the problem that we address, describes the approach and the data and metrics we use for its evaluation, and analyzes the experimental results guided by our research questions.

2 Methodology

2.1 Problem

Given an SDG target, and one or more passages from environmental impact assessment (EIA) reports, *SDG Target Evidence Identification* (or *SDG-EI* for short) is the task of deciding which of the passages, if any, is a relevant evidence supporting that the content of the target is addressed (Garigliotti, 2024). We assume an instantiation of this problem where a method, specifically an Large Language Model (LLM), is required to generate an answer –to a question asking for deciding which among the candidate passage(s) is relevant– such that the answer contains the correct passages, each referred to by a unique string identifier also included in the prompt used to generate with the LLM. Figure 1 presents an overview of our problem and illustrates it with examples of evidence identification for SDG targets.

2.2 Retrieval-Augmented Generation

We approach the SDG-EI problem by instantiating the Retrieval-Augmented Generation (RAG) framework in a series of methods. A main characteristic of RAG is its ability to complement the parametric or implicit information in an LLM with external, domain-specific knowledge provided in the prompt issued to the generator LLM (Lewis and et al., 2020; Gao et al., 2023). As its name indicates, RAG consists in three main stages. The first one, *retrieval*, obtains

a ranked list of documents for a query from a given document index. In our scenario, a collection of EIA reports is indexed and retrieved from by issuing the content of an SDG target as the query. In the second stage, a standard prompt –designed to ask an LLM to answer about relevant EIA passages for a given target– is *augmented* with the retrieved passages so that the generator is provided with these as context for the question. Otherwise, the SDG-EI task in it would be left to face the entirety of the space of implicit knowledge in the LLM where most likely the desired output for the passages of interest would not be found. The final phase involves inputting this augmented prompt into an LLM from which to *generate* an answer containing the desired output, in this case, the mention of the passages relevant to the SDG target, if any.

A template of the prompts developed in this work is shown in Table 1. The settings for the parameters studied here are described in Section 3.2.

2.3   Explainable SDG-EI

In order to operationalize the elicitation to a model for inspecting its own rationales used to produce the generated output, we directly request to the LLM to complement its answer about passage relevance to a target with an explanation. Specifically, we also add to the prompt, during augmentation, (i) an additional instruction, mentioning the intention to ask for this self-inspection, and (ii) a complementary question asking for explaining on the generated answer.

In the template presented in Table 1, the placeholders [XAI instr.] and [XAI Q.] are replaced by the actual explainability-oriented instruction and question, according to experimentation with relevant parameters described in Section 3.2.

3   Experimental Setup

3.1   Dataset

We make use of a collection based on environmental reports made publicly available by the Ministry of Climate of the Republic of Estonia.[2] These reports correspond to projects developed in the country and other European countries nearby, hence determining a particular environment with the geography, the human activities and regulations that exists in these locations. The collection comprises 16,474 passages obtained from 33 reports. It also collects the relevance of passages –obtained with a baseline lexical retrieval method– for a selection of 30 SDG targets, out of the 157 targets within the SDG framework. These targets are relevant to the environmental assessment carried out in said environment of provenance for the reports, and so more likely to be addressed in them. Figure 2 lists all the 30 selected SDG targets.

The dataset is included in our resources made publicly available.[3]

---

[2] `https://kliimaministeerium.ee/piiriulene-moju-hindamine#piiriulese-moju-hind`

[3] `https://bit.ly/AIMM_at_ER_2024-SDG_EI_XAI-Materials`

6.1, 6.3, 6.4, 6.6

7.1, 7.2, 7.3

9.1, 9.2, 9.4

11.1, 11.2, 11.3, 11.4

12.2, 12.3, 12.4, 12.5

13.1, 13.2

14.1, 14.2, 14.3, 14.4, 14.5

15.1, 15.2, 15.3, 15.5, 15.8

Fig. 2: The 30 SDG targets in the dataset.

3.2   RAG Configurations

In this work, each SDG-EI method is an instance of the RAG framework, where a particular assignment of values is set to our RAG parameters of interest, in what we refer to as a (parameter) configuration.

For the **retrieval** phase, we use the SDG target as query to retrieve the top 3 passages (i.e. 3 the cut-off) from the indexed collection of passages, via both traditional lexical retrieval (lexical, for short) and learned dense retrieval (dense, for short) (Gao et al., 2023). The well-established Pyserini library[4] is here used to perform retrieval.

In the second stage, **augmentation**, we experiment with the way that the retrieved passages are integrated. Within a common format where each passage is provided preceded by its unique identifier from the collection, a straightforward augmentation lists the passages in the same order as in the retrieval output. Since the presence of implicit artefacts was detected where the model tries to memorize the top positions of the relevant passages in the retrieval order, we also experiment with an alternative *ordering* where passages appear in the prompt in a random order. In this same stage we also experiment on few-shot learning, specifically, with the *number of examples* provided in the prompt to help narrow down the generation space of the LLM –the possible values are 1 and 2, i.e. one-shot and two-shot. A third parameter of interest is the actual *explanation question* that elicits the rationales from the LLM. We distinguish the basic prompt, i.e. without explanation elicitation, from the alternative ones that we refer to as XAI-aware prompts. A direct XAI question, *X-direct*, asks "Why

---

[4] https://github.com/castorini/pyserini

Table 1: Template to build the basic prompt during augmentation. The templates for the XAI-aware prompts are almost identical except for the enabled explainability components **[XAI instr.]** and **[XAI Q.]** omitted in the basic prompt.

Prompt template
You are an assistant for tasks in environmental impact assessment (EIA). A few excerpts from the textual content of EIA reports are provided by the user as contexts. Please ANSWER the QUESTION about the possible relevance of these contexts for the given Sustainable Development Goal (SDG) target. Please answer to the best of your ability. If you don't know the answer, just say that you don't know. Keep the answer concise. When you refer to a context in your answer, always cite the corresponding context ID (which must be among the given CONTEXTS) between square brackets (e.g. [a1b2x34d]), as it's done in each example. Examples are given below, each example between the '<example>' and '</example>' tags. After that, you are given the actual SDG target with contexts so that you answer about it. **[XAI instr.]**

Prompt template (ctd.)
⟨example⟩ ... ⟨/example⟩ ... QUESTION: Which one(s), if any, of the provided context(s) is a relevant evidence where the SDG target is addressed? **[XAI Q.]** CONTEXTS: Context ID: ... Context: ... ... ANSWER:

do you think that this is the answer to the question?" right after the question for passage relevance to the SDG target. A counterfactual strategy, *X-counter*, instead, inquires for alternative prompting scenarios challenging the augmented LLM input itself: "What would you have answered to the same question if the order of the passages in the prompt was different?" Each of these XAI-aware prompts correspond to the template in Table 1 where the respective question mentioned above is put in the [XAI Q.] placeholder. With the same template, in a XAI-aware prompt the placeholder [XAI Instr.] at the end of the prompt header is replaced by the additional instruction "You are also asked about why you are giving this answer to the question. Please respond to it right after."

   Generation is finally performed by inputting the prompt into the GPT-3.5 (gpt-3.5-turbo-0125) LLM (Radford et al., 2019).

3.3   Evaluation Metrics

We evaluate the SDG-EI performance of each of the proposed RAG methods by the means of standard set metrics of precision, recall and F-score with respect to the predicted and relevant passages for an SDG target. Given a method, we report the average performance across all its SDG targets.

4   Experimental Results

Table 2 presents the results for all our experimental configurations. Each block in the table corresponds to one of the prompting strategies: basic, X-direct, and

Table 2: Experimental results for all the configurations in the SDG Evidence Identification task. A metric group indicates the setting for the parameter about number of few-shot examples in the prompt (one or two). In all these experiments, the retrieval cut-off is top 3 results and the generator LLM is GPT-3.5. In each block of this table, the best performance on a metric is shown in **bold**.

Retrieval method	Passage order	One-shot			Two-shot		
		Prec.	Rec.	F-Sco.	Prec.	Rec.	F-Sco.
Prompt: Basic (no XAI elements)							
Lexical	By ranking	0.7222	**0.6611**	**0.6656**	0.7667	**0.6778**	0.6944
	Random	**0.7444**	0.6	0.6411	**0.7944**	0.6722	**0.6956**
Dense	By ranking	0.6556	0.6389	0.62	0.6556	0.5667	0.5878
	Random	0.6833	0.5833	0.5989	0.7	0.6056	0.6322
Prompt: X-direct							
Lexical	By ranking	**0.7889**	**0.8167**	**0.7689**	0.7889	**0.6611**	**0.6833**
	Random	0.7333	0.6	0.6111	0.7867	0.6389	0.6617
Dense	By ranking	0.6889	0.6111	0.5989	0.7667	0.5778	0.6156
	Random	0.6889	0.65	0.6456	0.7111	0.5389	0.58
Prompt: X-counter							
Lexical	By ranking	**0.8**	**0.4444**	**0.5489**	0.7667	0.3889	0.4956
	Random	0.7333	0.3444	0.46	**0.8333**	**0.4389**	**0.5511**
Dense	By ranking	0.7333	0.3722	0.4711	0.7	0.3722	0.46
	Random	0.7	0.3944	0.48	0.7667	0.4333	0.5222

X-counterfactual, resp. The corresponding files with the full results for every RAG stage are made available in the public directory with our resources.[5]

We analyze these experimental results by answering our research questions.

RQ1: *How do methods perform in terms of retrieval-stage and few-shot parameters?* In all the settings and for every metric, lexical retrieval clearly results to perform better. Qualitatively, these cases are often favoured by few words that are very relevant for correctly matching query –target– and passage during retrieval. These are words that, while distinctive for a target and/or passage, become less distinctive when combined with the semantics of other words by dense retrieval. Examples of these key words found in our data are "transport" (for target 11.2), "acidification" (for target 14.3), "overfishing" (strong signal for SDG target 14.4), and "alien" (for target 15.8 about invasive species).

RQ2: *What is the impact of the order of the passages in the prompt?* We observe strong indications for our hypothesis about the existence of artefacts in the LLM that make it favour the top passages. The methods where the passages

---

[5] https://bit.ly/AIMM_at_ER_2024-SDG_EI_XAI-Materials

are provided for augmentation in the same order as in the retrieval ranking outperform the counterparts with random order of passages in the prompt, i.e. the respective method with order by ranking gets affected when its passages do not keep the best on top. In the particular cases of X-counter prompt with two examples we observe the inverted pattern, and we suspect it has to do with an LLM confused by the counterfactual request.

RQ3: Does self-explanatory generation help improve the performance on SDG-EI? In both precision and recall metrics, and hence in their harmonic mean F-score too, SDG-EI is mostly benefited by X-direct as a simple explainable generation mechanism that we request in the prompt. X-counter, too, shows improvements for the best performing configurations in terms of precision. However, measuring by recall, X-counter clearly hurts the performance with large drops by the increased absence of correct passages in proportion in the answer. It seems then to confirm the confusion phenomena when inquired by counterfactual about passage ordering.

5   Conclusions and Future Work

In this paper, we study the usage of Retrieval-augmented Generation methods powered by an established commercial LLM for SDG evidence identification. We assess the fundamental ability of the LLM to generate the expected output when provided with explicit, external knowledge in a well-engineered prompt. Plus, we evaluate it in comparison with simple strategies to enable explainable generations by eliciting self-explanatory answers.

This evidence not only aims to confirm the prediction of a model about the content of a report addressing the SDG target in question, but also may serve to support related conceptual modeling tasks. Prominently, we envision addressing the mapping of a report –or particular parts of its content– into the structured counterparts in an ontology of environment-centered items such as activities, impacts and targets. The study of problems on linking mentions of these environmental elements to the uniquely identified entries in a structured knowledge resource is a main line of future work.

References

A. Asai, M. Gardner, and H. Hajishirzi. Evidentiality-guided generation for knowledge-intensive NLP tasks. In *Proceedings of NAACL-HLT*, pages 2226–2243. ACL, 2022.

E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, and N. Nobani. A survey on XAI and natural language explanations. *Information Processing & Management*, 60(1):103111, 2023. ISSN 0306-4573.

A. Del Campo, P. Gazzola, and V. Onyango. The mutualism of strategic environmental assessment and Sustainable Development Goals. *Environmental Impact Assessment Review*, 82:1–9, May 2020. ISSN 0195-9255.

Y. Elazar and A. B. et al. What's in my big data?, 2024.

T. Gao, H. Yen, J. Yu, and D. Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488. Association for Computational Linguistics, 2023.

D. Garigliotti. SDG target detection in environmental reports using retrieval-augmented generation with LLMs. In *Proceedings of the 1st ClimateNLP Workshop*, pages 241–250. ACL, 2024.

D. Garigliotti and J. B. et al. Do bridges dream of water pollutants? towards dreamskg, a knowledge graph to make digital access for sustainable environmental assessment come true. *Proceedings of the ACM Web Conference*, 2023.

P. Lewis and E. P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

N. Liu, T. Zhang, and P. Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025. ACL, 2023.

J. Menick and M. T. et al. Teaching language models to support answers with verified quotes. *ArXiv*, abs/2203.11147, 2022.

F. Å. Nielsen and I. L. et al. Environmental impact assessment reports in wikidata and a wikibase. In *ESWC Workshops*, 2023.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019.

D. Spokoyny, T. Laud, T. Corringham, and T. Berg-Kirkpatrick. Towards answering climate questionnaires from unstructured climate reports, 2023.

D. Stammbach and N. W. et al. Environmental claim detection. In *Proceedings of the ACL (Volume 2: Short Papers)*, pages 1051–1066. ACL, 2023.

D. Thulke and Y. G. et al. ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change. *ArXiv*, abs/2401.09646, 2024.

H. Touvron and L. M. et al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.

S. A. Vaghefi and D. S. et al. ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, 4, 2023. CRIS-Team Scopus Importer:2023-12-29.

N. Webersinke, M. Kraus, J. A. Bingler, and M. Leippold. ClimateBert: A Pretrained Language Model for Climate-Related Text. In *Proceedings of AAAI 2022 Fall Symposium*, 2022.