SDG target detection in environmental reports using Retrieval-augmented Generation with LLMs

Darío Garigliotti
University of Bergen
Norway
dario.garigliotti@uib.no

Abstract

With the consolidation of Large Language Models (LLM) as a dominant component in approaches for multiple linguistic tasks, the interest in these technologies has greatly increased within a variety of areas and domains. A particular scenario of information needs where to exploit these approaches is climate-aware NLP. Paradigmatically, the vast manual labour of inspecting long, heterogeneous documents to find environment-relevant expressions and claims suits well within a recently established Retrieval-augmented Generation (RAG) framework. In this paper, we tackle dual problems within environment analysis dealing with the common goal of detecting a Sustainable Developmental Goal (SDG) target being addressed in a textual passage of an environmental assessment report. We develop relevant test collections, and propose and evaluate a series of methods within the general RAG pipeline, in order to assess the current capabilities of LLMs for the tasks of SDG target evidence identification and SDG target detection.

1 Introduction

A series of Sustainable Development Goals (SDGs) were established by experts in the United Nations, as a reference framework with respect to which guide the progress of human activities, altogether oriented to the common good (Del Campo et al., 2020). According to their respective legal requirements, practitioners in the area of environmental assessment (e.g., professional assessors, developers, authorities) have to incorporate this framework in multiple spheres. In particular, the activities, impacts and mitigation measures described in environmental assessment documents are increasingly required to report how they address one or more SDGs; especially SDG targets, these being focused, actionable subgoals within a given SDG. Identifying textual passages relevant in addressing an SDG target of interest then becomes a fundamental problem in the practice of environmental assessment. This problem naturally lends itself within an approach based on obtaining an initial selection of passage candidates, and then apply onto this a more advanced detection technique to refine the decisions about relevance to the target. The recently established Retrieval-augmented Generation (RAG) framework (Lewis et al., 2020) embraces this basic approach and couples it with a text generation component powered by Large Language Models (LLMs), the dominant technology in NLP that shows state-of-the-art performance in multiple tasks (Radford et al., 2019; Touvron et al., 2023).

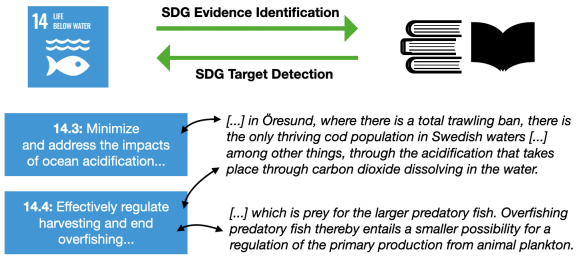

Figure 1: The dual tasks that we address in this work: SDG Evidence Identification and SDG Target Detection.

Figure 1 depicts an overview of the two tasks that we address in this work, with examples of SDG targets and excerpts of environmental impact assessment (EIA) reports. Specifically, these dual tasks couple with each other in the need of practitioners in EIA for finding instances of correspondence between the information spaces of SDG targets and textual evidence in specialized reports. We contribute by developing a test collection for each of these two tasks. We also propose and evaluate a series of experimental configurations for each of the RAG components, in order to assess the current capabilities of LLMs for these paradigmatic tasks in climate-aware NLP.

Our test collection and related resources developed in this study are made publicly available in a repository at https://bit.ly/

climatenlp-sdg-target-detection.

2 Related Work

Recently, the area of climate-aware natural language processing has decisively emerged led by a broad interest of developing information access methods to strengthen awareness in phenomena within climate change, as well as to process climate-related data in specific tasks within this domain. For example, Hershcovich et al. (2022) introduce a climate performance model card to summarize the impact of the experimentation corresponding to a scientific work in NLP research. Some of the related literature focuses on designing and analyzing methods for extracting climate-centric information, for example, to answer questionnaires (Spokoyny et al., 2023) and to detect climate-relevant claims in documents (Stammbach et al., 2023). Works like the ones by Bingler et al. (2022) and Schimanski et al. (2024) reveal the capabilities of well-established language models in communicating around climate awareness. Other lines of research in this area have produced fundamental resources, including language models such as ClimateBert (Webersinke et al., 2022) and ClimateGPT (Thulke et al., 2024), as well as systems like ChatClimate (Vaghefi et al., 2023) and resources like EIA-centric ontologies (Nielsen et al., 2023; Garigliotti et al., 2023), to power approaches for a variety of tasks.

An ever-increasing dominant technology in NLP, LLMs store vast amounts of information implicitly in their of billions of parameters trained on large general-purpose corpora, which allows them to perform as state of the art in many tasks such as text classification, textual entailment and question answering (QA) (Radford et al., 2019; Touvron et al., 2023). Yet, for many domain-specific scenarios, a framework like Retrieval-augmented Generation (RAG) (Lewis et al., 2020) becomes convenient as it allows to extend the LLM capabilities with additional explicit knowledge as context where to get the generated answer from. Moreover, against frequent hallucinations, it useful to be able to verify that the claims that occur in a text generated by such a model are truthful (Liu et al., 2023a; Menick et al., 2022). Within the research on QA, a foundational task for any application of LLMs, several related problems such as attribution in question answering (Bohnet et al., 2023), evidentiality-guided generation (Asai et al., 2022), verifiability of generation (Liu et al., 2023a), and factuality in summarization (Liu et al., 2023b) aim for operationalizing such a verification.

In the tasks we address, we treat the input (an SDG target in EI or an textual excerpt in TD) as a question for which, after augmented with retrieved items, an LLM must generate an answer with the correct outputs (identifiers of report passages or SDG targets, respectively) among the ones provided in the prompt.

3 Methodology

3.1 Problems

Given (i) an SDG target and (ii) a set of one or more passages from environmental impact assessment (EIA) reports, the task of *SDG Target Evidence Identification* (or *EI* task) consists in determining whether any of the passages is a relevant evidence where the content of the target is addressed. We instantiate this problem by requesting a method, specifically an Large Language Model (LLM), to generate an answer (to a question asking for deciding which passage(s) are relevant) with the correct passages, each referred to by a unique string identifier also provided in the generation prompt.

We assume the dual task of detecting SDG targets in a passage to be defined as follows. Given (i) a passage from an EIA report, and (ii) a set of one or more SDG targets, the task of *SDG Target Detection* (or *TD* task) consists in determining whether the content of any of the targets in the set is being addressed in the passage. Similar to EI setting, we request an LLM to generate an answer with the correct targets, each referred to by a unique string identifier provided in the prompt.

These two tasks together encompass an operationalization of typical dynamics in the practice of environmental analysis where a professional assessor aims to find correspondences between SDG targets of interest and textual evidence in reports.

3.2 Approach

We approach each of these two tasks with a series of methods within Retrieval-augmented Generation (RAG), a framework used, among others, by relevant benchmark literature for question answering (Gao et al., 2023). Each method consists of the same pipeline of three distinguished components: (i) retrieval, which, for each input question requested as query, produces a list of ranked passages from the indexed passage collection; (ii) then,

augmentation, where each test instance made of the question and (a subset of) retrieved passages are aggregated in a well-designed prompt that also captures the criteria of relevance to be required for judgement to an LLM; and (iii) a third component, the LLM-based generation of the answer required for each instance. Within this RAG umbrella framework, for example, for the problem of SDG evidence identification (EI), the SDG target becomes a query for which to identify relevant textual passages from a collection. Symmetrically, for the problem of SDG target detection (TD), a textual passage, or excerpt, is treated as a query for which to find the SDG targets, if any, among selected suitable target candidates; these targets are "passages" themselves within RAG since they are retrieved —from an indexed collection as candidates— for the input excerpt and subsequently post-processed through the augmentation and generation stages.

We experiment with corresponding parameters of interest on a vanilla setting of each component, and evaluate all the respective performances. We refer with 'method' to each instantiation of this RAG-based approach in a particular parameter configuration.

3.3 Research Questions

We conduct experimentation over the test collections with an ensemble of methods, in order to answer the following research questions.

- **RQ1**: How do the retrieval component affect the RAG performance?
- **RQ2**: What is the impact of the different augmentation strategies?
- **RQ3**: How does RAG perform with each LLM chosen for generation?

4 Experimental Setup

4.1 Datasets

The Ministry of Climate of the Republic of Estonia has made publicly available a series of environmental reports corresponding to projects developed in the country and other European countries nearby.[1] We select 33 reports from this public website and post-process their PDF files to obtain a collection of passages, or contexts, as follows. First, the textual content of each file is extracted with the PyMuPDF tool [2]. Then, further replacements in the text are made for distinguished characters so that to transform each sentence that is broken into multiple lines as appearing in the PDF, and recover each contiguous sentence. From these, we only retain every sentence made of at least 5 words; this allows to remove spurious content that is wrongly processed as a valid sentence. Finally, we select a random passage length (between 3 and 5 sentences) for each page, and chunk the content of that page into passages of that length, possibly with shorter trailing passages. Each passage is assigned a unique identifier, with respect to which is then indexed during the first stage of the RAG framework for the EI task. This identifier is the one requested to be in the answer generated by an LLM to refer to each passage that the LLM considers to be relevant to the given SDG target. We use unique random alphanumeric strings as identifiers aiming to avoid allowing that the LLM may hallucinate typical reference markers such as natural numbers [1], [2], etc. The obtained collection comprises 16,474 passages.[3]

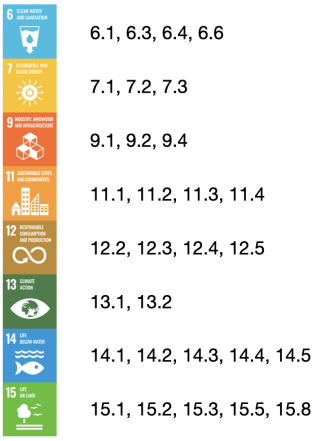

Figure 2: The 30 selected SDG targets in our datasets.

We also select 30 SDG targets, a considerable subset of the 157 targets available within the SDG framework. The selected targets are considered more relevant to the kind of environmental assessment practice of our interest, and so more likely to be addressed in them. Specifically, each target belongs to one of the following SDGs: Clean water and sanitation (SDG 6), Affordable and clean energy (SDG 7), Industry, innovation and infrastructure (SDG 9), Sustainable cities and communi-

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

[2] https://pymupdf.readthedocs.io/en/latest/

[3] The full list of links to the PDF documents for the reports, as well as the postprocessing of the files into the final passage collection, are made publicly available in our repository at https://bit.ly/climatenlp-sdg-target-detection.

ties (SDG 11), Responsible consumption and production (SDG 12), Climate action (SDG 13), Life below water (SDG 14), and Life on land (SDG 15). The full list of 30 selected SDG targets is presented in Fig. 2. After removing from it any of the temporal phrases like "by 2030" that are common to most SDG targets, the textual description of each of these selected targets becomes a pseudo-document, a "passage" by itself. The collection of these targets-as-passages is indexed, and retrieved against, during the RAG retrieval stage for the TD task. From the set of selected SDG targets and the passage collection, we obtain a two datasets, each per task, to evaluate the performance of our proposed methods. The test collection for the evidence identification task consists of manual annotations for the "yes"/"no" binary relevance judgement of a passage with respect to a target. The test collection comprises the 30 SDG targets, each annotated for 6 passages (the top 3 retrieved contexts for each of the two retrieval methods). Similarly, we build the test collection for the target detection task, by manually judging the binary relevance of a retrieved target w.r.t. an excerpt, for 10 "passages" (the top 5 retrieved targets for each of the two retrieval methods).

Both test collections are also made publicly available in our repository.[4]

4.2 Experimental Parameters

Retrieval component. For the EI task, from the index built to store the uniquely identified passages, we retrieve the top 3 results for every SDG target with each of both methods, traditional lexical matching (lexical, for short) and learned dense retrieval (dense, for short). For the TD task, we first build an index of SDG targets as passages, after each being assigned a random unique ID (this ID "masks" the plain number.subnumber ID format, as it is useful later for an augmentation configuration, where the LLM will not be made aware that the passages in the prompt are indeed SDG targets). We then retrieve top 5 targets from this index per each excerpt as query, again with both lexical and dense methods. We perform retrieval with the well-established library Pyserini.[5]

Augmentation component. Through prompt engineering, we design a prompt that requests the LLM to produce the answer mentioning the correct relevant passages in the desired format, which explicitly requires to give a concise answer and only if knowing it. Tables 1 and 2 show the actual prompt templates used for each task in one of our experimental configurations, 'SDG-explicit', where there is an explicit mention to the target being part of the SDG framework. The SDG-implicit prompt version is obtained from the explicit one by performing few replacements that mask an SDG target (as query in the EI task; as passage in the TD task) as being an environmental policy. For example, the SDG-implicit prompt for EI task is obtained from the prompt in Table 1 by replacing (i) "Sustainable Development Goal (SDG) target" by "environmental policy" in the prompt header, (ii) "SDG TARGET" by "ENVIRONMENTAL POLICY" in the field of the prompt footer where the SDG target is declared, and "SDG target" by "policy" in the question field by the end of the footer. The replacements to obtain a SDG-implicit prompt for the TD task are similar, with the additional detail of replacing each original target ID by its random unique identifier.

After observations about the phenomenon of an LLM possibly answering correctly most likely due

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. ⟨example⟩ ... ⟨/example⟩ ... Now, your task. CONTEXTS: Context ID: ... Context: ... ... SDG TARGET: ... QUESTION: Which one(s), if any, of the provided context(s) is a relevant evidence where the SDG target is addressed? ANSWER:

Table 1: Template to build the prompt during augmentation ('SDG-explicit' version) for the EI task.

[4] https://bit.ly/climatenlp-sdg-target-detection
[5] https://github.com/castorini/pyserini

Prompt template
You are an assistant for tasks in environmental impact assessment (EIA). An excerpt from the textual content of an EIA report is provided by the user. After it, 5 Sustainable Development Goal (SDG) targets are also provided, each target with its corresponding SDG target ID. Please ANSWER by identifying *all* the SDG targets that are relevant to be addressed in the context of the provided excerpt. 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 target in your answer, always cite the corresponding SDG target ID (which must be among the given SDG targets) between square brackets (e.g. [4.7]), 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 EIA excerpt so that you identify *all* the relevant SDG targets. ⟨example⟩ ... ⟨/example⟩ ... Now, your task. EXCERPT: ... SDG TARGETS: Target ID: ... Target: ... ... ANSWER:

Table 2: Template to build the prompt during augmentation ('SDG-explicit' version) for the TD task.

to learning the pattern about the passages in the prompt —being listed in the same order as the retrieved ranking—, we experiment with an alternative random order of contexts.

Generation component. We generate answers by prompting established LLMs. Specifically, we use a family of open LLMs such as Llama2 (Touvron et al., 2023) and a prominent model of the GPT platform, GPT3.5 (Radford et al., 2019). We also experiment with ClimateGPT, a family of LLMs obtained by fine-tuning a corresponding Llama2 model over corpora of documents within the climate change domain.

Generation with Llama2 and ClimateGPT is performed by inference with HuggingFace transformers library, while for GPT we access via the OpenAI API.

Summary. Our experimental parameters are:

- (Retrieval) Method: lexical or dense.
- (Augmentation) Prompt: SDG-explicit or SDG-implicit.
- (Augmentation) Number of examples: 1 or 2.
- (Augmentation) Order of passages: as given by the retrieval ranking, or random.
- (Generation) LLM: open (Llama2-13b, Llama2-13b-ch, ClimateGPT-13b) or closed (ChatGPT —gpt-3.5-turbo-0125—).

4.3 Evaluation Metrics

For each task, we evaluate the correctness of a method by applying standard retrieval metrics of precision and recall with respect to the retrieved passage set (all the passage identifiers mentioned in the generated answer) and the relevant passage set (the set of all the known relevant passages such that they appear among the contexts provided in the prompt). We remind that in the TD task, the SDG targets to be identified for a given EIA excerpt are considered to be the passages in the RAG framework. For a given method, we report the average performance across all the instances in the test collection of each task, i.e., across the 30 SDG targets for the EI task and across the 50 EIA excerpts for the TD task.

5 Results and analysis

Throughout this section, Tables 3 and 4 present the results for all the configurations in our experimentation. (The corresponding output files with the full RAG results for all methods are made publicly available in our repository.[6])

5.1 RQ1: Retrieval component

In our experimentation, the possible impacts of the retrieval stage are centered in the retrieval method: lexical or dense. Firstly, in the EI task, lexical retrieval leads to the best performances when combined with GPT3.5 or Llama2-13b-chat, across all metrics, and all augmentation strategies (number of examples and prompt version). Results with ClimateGPT-13b are split between the method setting, with more tendency to prefer dense retrieval, and mostly small changes across the parameter for the number of examples in prompt.

Secondly, in the TD task, we observe that, when using the SDG-explicit prompts, the precision measurements with ChatGPT are similar for a given setting and split for the number of examples, while its recall favours the dense retrieval method. Llama2-13b-chat also mostly changes between one- and two-example setting, regardless of the SDG-explicit or implicit prompt version. ClimateGPT-13b always performs best with lexical retrieval.

[6] https://bit.ly/climatenlp-sdg-target-detection

SDG-explicit prompt						
LLM	**Retrieval method**	**Passage order**	**One example**		**Two examples**	
			Precision	**Recall**	**Precision**	**Recall**
Llama2-13b	Lexical	By ranking	0.45	0.4444	0.1667	0.0778
		Random	0.4861	0.5333	0.1667	0.1
	Dense	By ranking	0.5556	**0.5944**	**0.3333**	**0.1833**
		Random	**0.5694**	0.5889	0.3167	0.1667
Llama2-13b-ch	Lexical	By ranking	**0.7**	0.5167	0.7556	0.65
		Random	0.6833	**0.5278**	**0.8056**	**0.7278**
	Dense	By ranking	0.6167	0.4444	0.6667	0.6
		Random	0.5333	0.4	0.6	0.5778
ClimateGPT-13b	Lexical	By ranking	0.6889	0.5333	0.6278	0.4889
		Random	0.6556	0.4722	**0.6833**	**0.55**
	Dense	By ranking	**0.7167**	**0.5944**	0.5833	0.5222
		Random	0.5611	0.5222	0.5833	0.5167
GPT-3.5	Lexical	By ranking	0.7222	**0.6611**	0.7667	**0.6778**
		Random	**0.7444**	0.6	**0.7944**	0.6722
	Dense	By ranking	0.6556	0.6389	0.6556	0.5667
		Random	0.6833	0.5833	0.7	0.6056

SDG-implicit prompt						
LLM	**Retrieval method**	**Passage order**	**One example**		**Two examples**	
			Precision	**Recall**	**Precision**	**Recall**
Llama2-13b	Lexical	By ranking	0.4778	**0.55**	0.2	0.0944
		Random	0.4611	0.4444	0.2333	**0.1556**
	Dense	By ranking	**0.525**	0.5389	**0.2667**	0.1389
		Random	0.4667	0.4389	0.3	0.15
Llama2-13b-ch	Lexical	By ranking	**0.7667**	0.5444	**0.7556**	**0.6222**
		Random	0.75	**0.55**	0.7389	0.6
	Dense	By ranking	0.6333	0.4611	0.6333	0.5444
		Random	0.6333	0.4389	0.5833	0.5
ClimateGPT-13b	Lexical	By ranking	**0.6889**	**0.5556**	0.5722	0.4056
		Random	0.6722	0.5056	0.5611	0.4389
	Dense	By ranking	0.6056	0.5167	**0.5833**	**0.4889**
		Random	0.5833	**0.5556**	0.55	0.45
GPT-3.5	Lexical	By ranking	0.7667	**0.6722**	**0.7556**	**0.7167**
		Random	**0.7833**	0.6056	0.7444	0.6556
	Dense	By ranking	0.6056	0.5889	0.65	0.5889
		Random	0.65	0.5611	0.6667	0.5722

Table 3: Experimental results for all the configurations in the SDG Evidence Identification task (*SDG-explicit* prompt version in the top half of the table; *SDG-implicit* prompt in the bottom half). A metric group indicates the setting for the parameter about number of examples in the prompt (one or two). For a given metric, the best performance on each LLM is in **bold** and the best overall performance is underlined.

The scenarios where lexical retrieval is favoured are possibly favoured by few key words that boost the correct matching during retrieval as they are very distinctive for a target and/or passage, which gets less distinctive when combined by dense retrieval with the semantics of other words. Examples of these key words found in our data are "overfishing" (strong signal for SDG target 14.4),

SDG-explicit prompt						
LLM	**Retrieval method**	**Passage order**	**One example**		**Two examples**	
			Precision	**Recall**	**Precision**	**Recall**
Llama2-13b	Lexical	By ranking	**0.437**	0.5933	**0.044**	**0.06**
		Random	0.3	0.416	0.01	0.02
	Dense	By ranking	0.4347	**0.609**	0.02	0.04
		Random	0.3313	0.4743	0	0.0
Llama2-13b-chat	Lexical	By ranking	0.38	0.58	0.423	0.5877
		Random	**0.396**	**0.6133**	0.332	0.447
	Dense	By ranking	0.294	0.491	**0.4677**	0.63
		Random	0.2253	0.365	0.4675	**0.676**
ClimateGPT-13b	Lexical	By ranking	0.6563	**1.0**	0.652	**1.0**
		Random	**0.6603**	0.9893	**0.662**	0.9793
	Dense	By ranking	0.611	0.98	0.608	0.98
		Random	0.6117	0.9433	0.6213	0.976
GPT-3.5	Lexical	By ranking	0.8783	0.6193	0.87	0.609
		Random	**0.8867**	0.602	0.8667	0.5827
	Dense	By ranking	0.8683	0.6893	**0.89**	**0.727**
		Random	0.857	**0.6927**	0.86	0.6887

SDG-implicit prompt						
LLM	**Retrieval method**	**Passage order**	**One example**		**Two examples**	
			Precision	**Recall**	**Precision**	**Recall**
Llama2-13b	Lexical	By ranking	**0.476**	**0.6127**	**0.06**	**0.045**
		Random	0.3957	0.4737	0.0	0.0
	Dense	By ranking	0.414	0.5393	0.015	0.02
		Random	0.3007	0.3413	0.0	0.0
Llama2-13b-chat	Lexical	By ranking	0.6583	0.9633	0.3677	0.3837
		Random	**0.6857**	0.8717	0.2823	0.286
	Dense	By ranking	0.6283	**0.975**	**0.3763**	**0.4483**
		Random	0.6287	0.8343	0.2667	0.3177
ClimateGPT-13b	Lexical	By ranking	0.649	**0.958**	0.654	**0.983**
		Random	**0.6587**	0.8697	**0.7067**	0.917
	Dense	By ranking	0.608	0.9467	0.613	0.975
		Random	0.6013	0.8327	0.627	0.8843
GPT-3.5	Lexical	By ranking	**0.93**	0.5893	**0.91**	0.527
		Random	0.89	0.5657	**0.91**	0.5387
	Dense	By ranking	0.86	0.5907	0.87	**0.5897**
		Random	0.8883	**0.5937**	0.8167	0.549

Table 4: Experimental results for all the configurations in the SDG Target Detection task (*SDG-explicit* prompt version in the top half of the table; *SDG-implicit* prompt in the bottom half). A metric group indicates the setting for the parameter about number of examples in the prompt (one or two). For a given metric, the best performance on each LLM is in **bold** and the best overall performance is underlined.

"acidification" (for target 14.3), "transport" (for target 11.2), "alien" (for target 15.8 about invasive species).

In the SDG target detection task, we observe that the the most frequent relevant targets belong to SDG 14 (about marine protection), which makes sense as most of the base reports where passages are taken describe aspects of environments in re-

gions around the Baltic Sea. It is followed in frequency by SDG 15 (terrestrial and inland freshwater ecosystems, forests), and with clearly less frequency by SDGs 7 (energy), 9 (infrastructure), 11 (housing, transportation), and 12 (waste, resources).

5.2 RQ2: Augmentation component

Results from the ablation of the augmentation component can be summarized as follows. Firstly, we analyze the impact of the order of the passages in the prompt. For the EI task, the order varies a lot w.r.t. other parameter settings in the SDG-explicit prompt configurations, whereas with the implicit prompt version, most cases favour the order by ranking. For the TD task, variations with SDG-explicit prompt version persist, while it varies slighlty less with implicit prompt and in many cases favouring order by ranking.

Secondly, we discuss the influence of the number of examples in the prompt. In the EI task, having two examples is mostly beneficial for Llama2-13b-chat and ChatGPT, while it harms ClimateGPT-13b performances and largely hurts Llama2-13b. In the TD task, the trends are similar for the Llama2 models but for ClimateGPT-13b and ChatGPT the results are mixed, with cases of clear disadvantage with more examples in the prompt.

5.3 RQ3: Generation component

Across both tasks and their respective configurations, we verify as expected that ChatGPT is the best performing LLM in several settings. A general pattern for the EI task is that GPT performs best in both metrics when only one example is provided in the prompt, followed by Llama2-13b-chat; and that this gets inverted as Llama2-13b-chat is the best performing in the two-examples setting. The base model Llama-13b performs very close to ClimateGPT-13b in very few scenarios, but the differences become clearer in favour of ClimateGPT-13b in the configurations with two examples.

For the TD task, GPT3.5 is the best performing LLM for both SDG-explicit and implicit prompt versions in the precision measurements. In turn, ClimateGPT-13b dominates in recall and clearly over Llama2-13b-chat for SDG-explicit prompt, but splits the best recall with Llama2-13b-chat in SDG-implicit, between one- or two-example settings, with Llama2-13b-chat overall closer.

5.4 Summary of observations

As a conclusive reiteration of our observations, we mention the following main remarks. (1) The EI task is best addressed with ChatGPT prompted with contexts obtained via lexical retrieval. (2) The TD tasks gets best precision-oriented performance when using ChatGPT over densely retrieved passages, while for best recall, it does with ClimateGPT over lexically retrieved passages. (3) In both tasks, most often the ranking in which passages where retrieved is the same order in which to list the passages in the prompt during augmentation. (4) The exact convenient number of examples in few-shot generation vary due to the complexity of the notion of a passage *addressing* an SDG target, and depends on the actual example(s) being considered.

6 Conclusions, Limitations, and Future Work

In this work, we study two dual problems on environmental analysis as a mean to approach towards the automatization of knowledge-intensive, time-consuming tasks in the practice of assessing environmental impact in reports and its correspondence with the recent developments around SDGs. Specifically, we propose and assess several methods within the RAG framework powered by LLMs.

Our work approaches a paradigmatic scenario of environmental analysis, yet it is still limited in its capabilities to identify evidence and detect targets. On the one hand, the selected targets cover a meaningful part of the SDGs scope in regards to EIA, yet there are more SDGs and targets that could be considered. On the other hand, the collection of reports where the EIA passages come from suits well as information source for our experimentation, yet it is centered on particular regions of Europe and so our study fails to capture phenomena about other environments and their corresponding SDG targets of relevance. Furthermore, our data annotation is conducted with caution and good faith but it could present cases where the judgement could be different, especially as the concept of "addressing an SDG (target)" is already not exact in the literature and the EIA practices described in the reports often take advantage of these uncertainties.

In future work, we plan to further study the duality of these two tasks by approaching environmental analysis with a method were each task retrofits the other one. In this way, for example, a textual

passage identified via EI for an SDG target can be the input of a subsequent TD stage to possibly expand the space of targets of interest for that EIA report, as well as exploiting relations between passages in the same report.

Another line of research is experimenting with the usage of a claim detector, this is, a dedicated model for identifying climate-aware claims in text, such as the one developed by (Stammbach et al., 2023). This component could complement the retrieval stage to improve the selection of passages that are finally fed into the LLM during generation.

A third possible area of work corresponds to automatically labeling larger volumes of test instances with an LLM as assessor, which could extend the evaluation space, as well as allow for experimenting with fine-tuning a base pre-trained model with these instances. In a similar fashion, a fourth direction would investigate the automatic assessment, also via LLM, of correctness for a predicted result. Such an assessment would be validated by observing the inter-annotator agreement with manual assessments in a sample of the test collection.

Acknowledgments

This paper is part of NRF project 329745 Machine Teaching For XAI.

References

Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2022. Evidentiality-guided generation for knowledge-intensive NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2226–2243, Seattle, United States. Association for Computational Linguistics.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures. *Finance Research Letters*, 47:102776.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and Modeling for Attributed Large Language Models. *Preprint*, arXiv:2212.08037.

Ainhoa González Del Campo, Paola Gazzola, and Vincent Onyango. 2020. The mutualism of strategic environmental assessment and sustainable development goals. *Environmental Impact Assessment Review*, 82:1–9.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. 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, Singapore. Association for Computational Linguistics.

Darío Garigliotti, Johannes Bjerva, Finn Årup Nielsen, Annika Butzbach, Ivar Lyhne, Lone Kørnøv, and Katja Hose. 2023. Do bridges dream of water pollutants? towards dreamskg, a knowledge graph to make digital access for sustainable environmental assessment come true. *Companion Proceedings of the ACM Web Conference 2023*.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023b. On improving summarization factual consistency from natural language feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nathan McAleese. 2022. Teaching language models to support answers with verified quotes. *ArXiv*, abs/2203.11147.

Finn Årup Nielsen, Ivar Lyhne, Darío Garigliotti, Annika Butzbach, Emilia Ravn Boess, Katja Hose, and Lone Kørnøv. 2023. Environmental impact assessment reports in wikidata and a wikibase. In *ESWC Workshops*.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 61. CRIS-Team Scopus Importer:2024-01-26.

Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. Towards answering climate questionnaires from unstructured climate reports. *Preprint*, arXiv:2301.04253.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *ArXiv*, abs/2401.09646.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, 4. CRIS-Team Scopus Importer:2023-12-29.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A Pretrained Language Model for Climate-Related Text. In *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*.