ULA, a Bibliometric Method to Identify Sustainable Development Goals using Large Language Models

Loris Bergeron
SnT - SEDAN
Banque de Luxembourg
Luxembourg, Luxembourg
loris.bergeron@bdl.lu

Jérôme François
SnT - SEDAN
University of Luxembourg
Luxembourg, Luxembourg
jerome.francois@uni.lu

Radu State
SnT - SEDAN
University of Luxembourg
Luxembourg, Luxembourg
radu.state@uni.lu

Jean Hilger
SnT - Finnovation Hub
University of Luxembourg
Luxembourg, Luxembourg
jean.hilger@uni.lu

***Abstract*—United Nations defined a set of 17 Sustainable Development Goals (SDGs) that must be derived by all states into concrete actions. As a result, methods need to be defined to evaluate the progress towards achieving those goals.**

However, evaluating each individual action with accurate measurements is not possible. As a result, many methods rely on analyzing textual documentation such as reports or publications to identify and comprehend the contributions of an entity to the different SDGs. Existing solutions are based on queries composed of a mostly manually fixed set of keywords. The exhaustiveness of these queries is strongly linked to the datasets used to build them but also to the personal interpretations of the SDGs.

To remedy this situation, we propose to extend a set of initial and manually validated keywords thanks to three major Large Language Models in order to generate and aggregate synonyms. For validation purposes, we rely on the OSDG Community Dataset which contains labelled text extracts alongside with the associated SDGs.

***Index Terms*—Sustainable Developement Goals (SDGs), United Nations (UN), Large Language Models (LLMs), GPT3.5, PaLM 2, Llama 2, Bibliometrics**

I. Introduction

In 2015, the agenda dedicated to sustainable development was adopted by all 193 member states of the United Nations (UN). At the core of the latter, a set of 17 Sustainable Development Goals (SDGs) are defined in [1]. As shown in table I, these objectives form a common plan to align all member states through a global partnership which aim to solve the major problems facing our world on different perspectives including fighting poverty or climate change for example. In addition to these 17 objectives, the UN have also defined 169 targets and 232 indicators associated with all the objectives, breaking them down into more granular levels. Targets describe the specific, measurable and achievable actions. For example, under SDG 1 (Eradicate poverty in all its forms everywhere in the world), target 1.1 aims *to eradicate extreme poverty completely by 2030*. To measure the progress towards each target, indicators have been defined. These are quantitative and qualitative means of assessing and analysing the level of achievement of a specific target. For example, one of the indicators for target 1.1 is the *proportion of the population living below the internationally agreed poverty line, by sex, age, employment status and place of residence*. The expressiveness of the definition of the various SDGs with this granularity is both a strength and a weakness. Although, the precise segmentation of the SDGs makes possible to clearly plan and execute concrete actions, it creates complexity in measuring their progress.

Publication produced in partnership with Banque de Luxembourg S.A.

TABLE I
The 17 Sustainable Development Goals (SDGs)

SDG	*Description*
SDG1	No Poverty
SDG2	Zero Hunger
SDG3	Good Health and Well-being
SDG4	Quality Education
SDG5	Gender Equality
SDG6	Clean Water and Sanitation
SDG7	Affordable and Clean Energy
SDG8	Decent Work and Economic Growth
SDG9	Industry, Innovation and Infrastructure
SDG10	Reduced Inequalities
SDG11	Sustainable Cities and Communities
SDG12	Responsible Consumption and Production
SDG13	Climate Action
SDG14	Life Below Water
SDG15	Life on Land
SDG16	Peace, Justice and Strong Institutions
SDG17	Partnerships for the Goals

When an entity (*e.g.* company, organization, country) takes into account the SDGs while deciding on related actions, they are usually advertised or reported through multiple channels including their various reporting (annual report, white papers, etc.). From a scientific perspective, the content of publications reflect the topics addressed by researchers or their affiliated organizations. Therefore, analyzing scientific publications have been advocated in the past to determine how they are aligned with and contribute to the SDGs [2]. The vast majority of techniques rely on creating a set of queries composed of fixed keywords to be applied on bibliographic databases, mainly titles and abstracts. In some cases, manual validation of the the keywords is necessary, thereby subjecting the work and the final results to a personal interpretation of the scope and definition of an SDG. The diversity of the keywords selected also depends on the nature, specificity, origin and diversity of the texts used to carry out the text analysis.

However, as we will show in our analysis, less than

half (41.16%) of text extracts related to a particular SDG are actually correctly distinguished with a recently proposed method [3]. This is due to the limited number of manually-defined keyword-based queries. In this paper, we are thus interested to investigate how an automated generation of queries can be helpful to improve this bibliography-based analysis.

The significant advances made in recent years in Generative Artificial Intelligence and more particularly those linked to Large Language Models (LLMs), open new opportunities for bibliographic approaches. These LLMs refer to a class of Artificial Intelligence (AI) designed to understand and generate human language in a sophisticated way. Unlike traditional models, they have been trained on vast textual datasets and contain tens of billions of parameters. As a result, they are capable of interpreting input data and produce appropriate output in natural language.

In this paper, we promote the use of LLMs to strengthen the mapping of a text extract to the SDGs it relates to. This is based by leveraging the generative capabilities of the LLMs to extend automatically the query set and complement the rigidity of original keywords defined in related works.

The contributions of this paper are three-fold:

- A new method to map text extracts to SDGs based on LLMs
- An implementation of the aforementioned method with three models (GPT-3.5, Llama 2, PaLM 2)
- A comparative result analysis of our method with one of the most recent state-of-the-art approach [3]

The rest of the document is structured as follows. Section II reviews related work. Our method is described in section III. Section IV focused on the experimental setup and the obtained results are reported in section V. Finally, section VI draws up an assessment of the research carried out and possible improvements for future work.

II. Related work

Several works proposes techniques to map scientific documents with SDGs.

In 2017, the Sustainable Development Solutions Network (SDSN) proposed a compiled list of 915 keywords[1], covering the 17 SDGs. It is the result of a combination of work from different universities to establish how their respective activities contribute to the SDGs: the Auckland University of Technology, Macquarie University, Monash University, Victoria University of Wellington, The University of Auckland, and the University of Western Australia. During the webinar organised by the SDSN[2], two main limitations have been highlighted: the time to manually select and agree on keywords and the lack of relevant keywords. In 2018, the Aurora project is building on the preliminary work of the SDSN by using the terms appearing in the SDG targets and indicators defined by the UN to form a first version of its own research [4] (v1.0). Various improvements to this work were carried out by Aurora until it reached version v5.0.2 [5] in 2020. The list of keywords drawn up for this latest version is based on a range of text analysis works as well as on the results of a survey [6] of experts in SDG-related areas of research. In 2019, Elsevier produced a list of 16 keyword-based queries [7] to enable researchers and institutions, through Scopus, to track and demonstrate the progress of their activities towards achieving the SDGs. This was followed by a number of improvements to the keywords, based on the documents returned by Scopus during the execution of the queries. In 2021, Elservier published a new work [8] improving the previously constructed queries by including those formulated by Aurora and combining them with the results of a logistic regression model. However, the human intervention of bibliometricians and analysts is necessary to establish the subsets per SDG and to construct the queries. Finally, the final validation of these queries is subject to validation by an expert. In 2023, the University of Auckland introduces the *Auckland Approach* (AA) [3] . It relies on the efforts of Elsevier, the SDSN and the UN, and uses analysis techniques to extract SDG-related keywords from publication metadata (*e.g.* title, keywords, and abstracts). Once the analysis methods have been applied, the most frequently occurring keywords are manually reviewed in order to refine them. The selected keywords are combined with those adopted by Elsevier, SDSN and the UN SDG indicators to form the final keyword lists *SDG Keywords Mapping*[3].

As highlighted, all these approaches still require a qualitative manual processing even when partially relying on automated analysis. Our method is based on the AA by automatically extending at a larger scale the proposed keywords using LLMs (up to 1138 more keywords).

III. Method

A. ULA overview

The overall process of our solution consists of multiple stages summarized in Figure 1 and detailed in sections hereafter. The initial set of keywords provided by AA are fed into the LLMs to generate synonyms. However, as the generation by such kind of Natural Language Processing (NLP) methods may produce formatting errors or generic keywords, this is followed by two stages of data cleaning, respectively stage 1 for performing corrections and stage 2 for filtering. This process ensures having high quality keywords. All of them are then aggregated with the initial keyword list (also filtered through stage 1) to create a refined and larger set. Finally, when a text extract must be analyzed to identify its associated SDG, the related keywords are queried into this text extract. It is worth mentioning that the text extract is also curated through a similar data cleaning process but this step is omitted in the figure for sake of clarity.

[1]List of SDG keywords compiled by SDSN

[2]Practical approaches to mapping university contributions to the Sustainable Development Goals (SDGs)

[3]The University of Auckland SDG Keywords Mapping

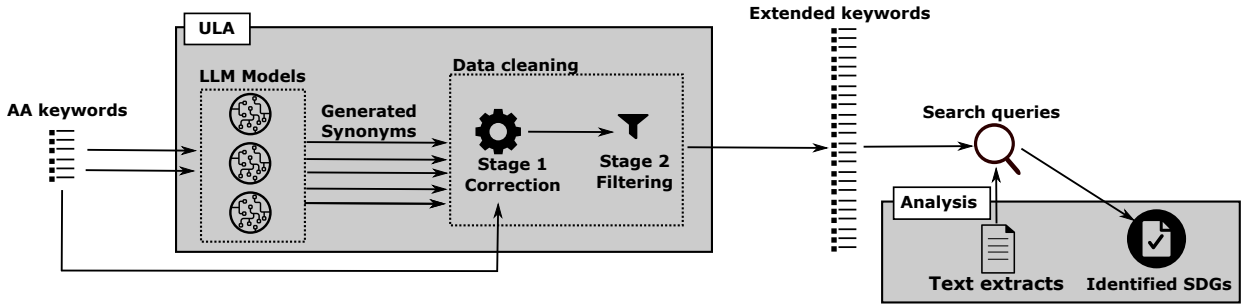

Fig. 1. ULA overview

B. Background on LLMs

A LLM is a large architecture of neural networks which can be trained on a vast amount of training data in order to support NLP tasks. Recent proposals mostly rely on the widely adopted transformer architecture leveraging self-attention mechanism [9].

A multitude of LLMs have become available in recent years. We have selected 3 LLMs considering those developed by 3 major companies heavily devoting part of their activities to AI:

- OpenAI - GPT-3.5 [10] (at the time of this research work version 4 was not yet available via the official OpenAI API): it is a set of models capable to understand and generate natural language or code. The model chosen for this paper is *gpt-3.5-turbo* version, which has been optimised for conversational tasks but also works for traditional completion tasks. It is based on training data till in September 2021 on a set of 175 billion parameters.
- Meta - Llama 2 [11]: this solution is positioned as a direct competitor to GPT. This second version was publicly released in July 2023 and has been trained with 40% much data than version 1. It offers different settings with 7, 13 or 70 billions parameters. A fine-tuned version of Llama 2, called Llama-2-chat, has been designed for conversational tasks. This fine-tuned version is based on publicly available training data sets and over a million human annotations. We chose the fine-tuned version of Llama 2 with 70 billion parameters to carry out our experiments.
- Google - PaLM 2 [12]: this represents Google's latest advances in reasoning tasks, including on programming code and mathematics, classification, question answering and even translation and generation of text in natural language. We selected it because of its ability to detect the nuances of human language and understand the ambiguous and figurative meanings of words, rather than their literal meaning. We used the *text-bison-001* variant as it is suited to linguistic tasks such as feeling analysis, answering questions or even rewriting text in a different style.

C. Keywords generation

In this paper, we consider different approaches to generate keywords $a \in A = \{AA, GPT, Llama, PaLM, ULA\}$ with AA the *Auckland Approach* [3] and the others, the LLM-based methods we propose and describe hereafter.

We denote as $SDG_i^a, a \in A$ and $i \in 1, \ldots, 16$ the set of keywords produced by the approach a for the SDG i. It is worth mentioning that the 17th SDG is not considered in this paper as it does not having associated query in the AA, which serves as the baseline.

Assuming the set of initial keywords from AA, we rely on the LLM models to generate synonyms. Especially, we conditioned them to behave like an English-language translator before asking them to generate s synonyms while being as explicit as possible, without generating explanations, formatting, acronyms or abbreviations. They were all conditioned in the same way thanks to the following prompt :

> *I'm asking you to play the role of English translator. I'm going to give you a text and I want you to generate 3 different synonyms. I don't want any explanations, formatting or even acronyms or abbreviations, just a text that everyone can understand. I want the synonyms I require and nothing else*

$syn_s^a(k)$ are the s generated synonym for the keyword *k* using the approach $a \in A - \{AA, ULA\}$, (*i.e.* a LLM model).

This results in defining the following set of keywords generated by an LLM as all generated synonyms from the original keywords of AA:

$$SDG_i^a = \bigcup_{k \in SDG_i^{AA}} syn_s^a(k), a \in A - \{AA, ULA\} \quad (1)$$

At the end of the different generations, we have 4 independent sets of keywords each composed of its own keywords. To remain as neutral as possible, no change was applied to the default configurations of LLMs (*e.g.* temperature, top_p).

Finally, the *University of Luxembourg Approach* (ULA) aggregates the four sets (original and extended keywords):

$$SDG_i^{ULA} = SDG_i^{AA} \cup SDG_i^{GPT} \cup SDG_i^{Llama} \cup SDG_i^{PaLM} \quad (2)$$

D. Post-generation data cleaning

The data cleaning process is applied to each set of keywords $SDG_i a$ and aims at ensuring that keywords are usable in search query afterwards and not too much generic.

In the first data cleaning stage, *DC1*, the process is composed of the following tasks: text conversion to resolve potential Unicode conversion issues, full wording of English language contractions, various steps to clean up the text (e.g. deleting spaces at the beginning and end of sentences, deleting line break characters), removing stop words or applying a stemming process

This first stage does not reduce the number of keywords. The second cleaning stage, *DC2*, aims at excluding keywords considered irrelevant due to their high genericity. For example, a unigram can be complex to interpret and a source of confusion in a bibliometric approach (*e.g.* the keyword *Hunger* can be at least linked to SDG1, SDG2, SDG3 and SDG10). All unigrams have been thus discarded. Besides, only keywords composed of at least two words are kept.

After being processed through *DC1* and *DC2*, each $SDG_i a$ is transformed into a final set $SDG_i^* a$.

IV. Experimental setup

A. Validation set

Using our 5 definitive sets of keywords, our aim was to compare their capabilities to identify SDG-related content in text extracts.

For a fair assessment of our approach, our validation is based on the OSDGD Community Dataset (OSDG-CD) [13], a labelled dataset of text extracts. Its latest version (as of July 2023) is made up of 42065 text extracts drawn from publicly available documents, as well as from a significant number UN-related sources such as SDG-Pathfinder and SDG Library. This dataset was labelled manually by bringing together world-wide volunteers such as researchers, subject-matter experts or SDG advocates. Volunteers taking part in this validation exercise must use their respective knowledge to validate or reject the correspondence between the available text extracts and a proposed relevant SDG. Through the exploratory analysis of the dataset, we noticed disparities in the volume of documents collected. For example, SDG16 is associated with 5451 documents unlike SDG 12 with 1108 related documents.

For each text extract and associated SDGs, the dataset provides the number of volunteers who agree and disagree on the proposed association. When a majority of volunteers disagree, we consider the text extract to not be associated with the given SDG. This filter has removed 7642 text extracts, leaving only 34423 in our final validation set. We also checked that there were no text extract present in more than a single SDG and we checked that there were no duplicates in the same SDG.

To sequence our work, we separated our 34423 text extracts into 16 subsets divided by SDGs. We conducted our comparative studies on these subsets by applying the appropriate queries on the *text* column. To make our results comparable we have applied the *DC1* process to the *text* column, so that we can apply the bibliometric queries in a uniform and consistent way.

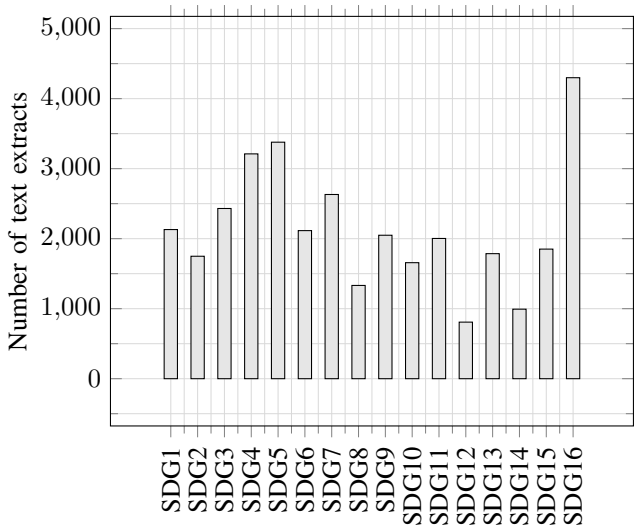

Fig. 2. Distribution of text extracts in our validation set

B. Creation and execution of queries

In order to generate queries to document as described in Figure 1, we set $s = 3$ as the number of synonyms to be generated by each LLM.

To give a comprehensive analysis of the LLM capabilities, we consider each individual set (AA, GPT, Llama, PaLM) in addition to the ULA knowing that AA is the baseline approach from the state-of-the-art.

A query consists in requesting all possibles keywords of a SDG using the *OR* operators and is applied on the text extracts of the OSDG-CD dataset (column named *text*). Hence a text extract is considered related to a given SDG if it matches at least one of the corresponding keywords. Thanks to the checks we carried out in section IV-A, our problem is a multi-class classification problem. Standards metrics are used such as precision, recall and F1-score alongside with a one-vs-all evaluation.

V. Results

A. Exhaustiveness of keywords

As a preliminary experiment, we were interested in assessing how much each approach can extend the original set of keywords to compare the generative capabilities of the underlying LLMs, knowing that generated keywords are cleaned and filtered. For example, if two generated synonyms use the same radix, they will be converted to the same keyword. Hence, even if several different synonyms are generated, the cleaning stages can end-up with less keywords.

At a global level, we considered the total number of keywords contained or generated by each approach:

$$|a| = |\bigcup_{i=1\ldots16} SDG_i a|$$

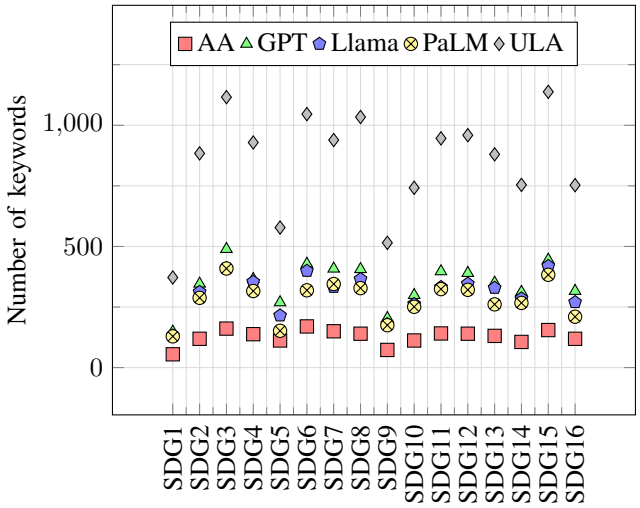

Fig. 3. Number of keywords after data cleaning

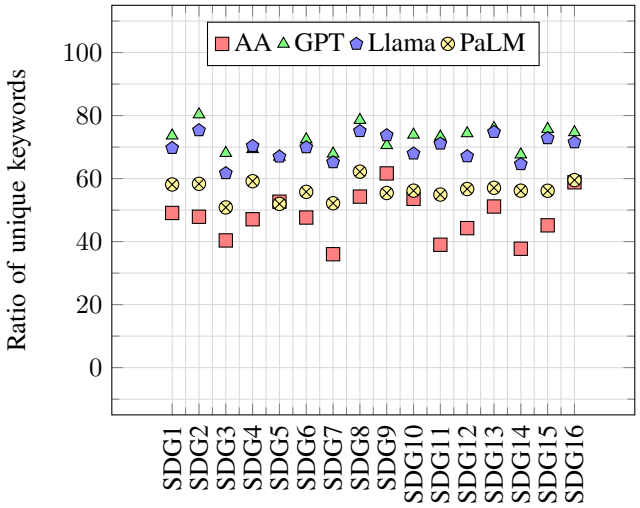

Fig. 4. Percentage of unique keywords ($100 * E_a^i$)

We obtained the following result: $|AA| = 2022$, $|GPT| = 5572$, $|Llama| = 4942$, $|PaLM| = 4478$ and $|ULA| = 13585$.

Figure 3 shows a sharp increase in the number of keywords contained in ULA compared with AA. In terms of the total number of keywords per SDG, AA shows a degree of stability, whereas ULA exhibits a greater disparity between SDGs. Nevertheless, we observe an overall similar increase with any LLM regarding the number of keywords. As an example, for SDG15, AA provides 155 keywords whereas ULA provides 1138 (*i.e.* an increase of 634%).

This highlights the complementary of creating queries that combine, on the one hand, the keywords drawn up in the classic way and, on the other hand, the keywords generated using LLMs.

Next results are focused on uniquely generated keywords by each LLM. We have evaluated the following ratio:

$$E_a^i = \frac{|SDG_i^a - \bigcup_{x \in A-\{a\}} SDG_i^x|}{|SDG_i^a|} a \in A \qquad (3)$$

This equation is not relevant for ULA as it represents an aggregation of all approaches. In Figure 4, GPT results in the higher exhaustiveness of keywords, followed by Llama and finally PaLM noting that PaLM is always the most under-performing LLM. AA also has the lowest ratio in 13 out of 16 cases.

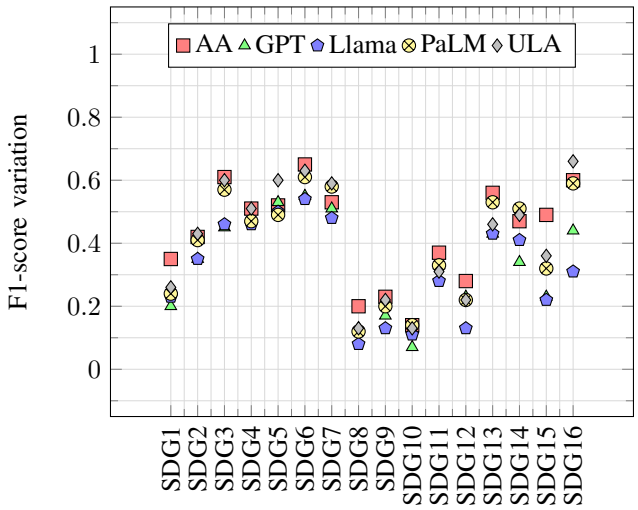

Fig. 5. F1-scores of the different approaches

TABLE II
MACRO-AVERAGED CLASSIFICATION METRICS

Approach	F1-score	Precision	Recall
AA	0.43	0.50	0.41
GPT	0.34	0.39	0.32
Llama	0.32	0.38	0.32
PaLM	0.40	0.40	0.43
ULA	0.41	0.36	0.55

B. Validity of generated keywords

Due to a significant increase in the exhaustiveness of keywords, assessing the positive and/or negative impact on the retrieved text extracts is performed on the validation set.

Figure 5 compares the different approaches based on the F1-score. As shown, the original AA approach can perform well on given SDGs due to its high specificity. This approach leads to the higher macro-averaged precision in table II. This table highlights (1) a significant room for improvement assuming the original AA and (2) strong disparities in performance between the LLMs. GPT and Llama significantly degrade the precision and recall and consequently the F1-score. However, the aggregated ULA model has an equivalent F1-score in comparison with AA (degradation of about 2%) with a higher recall (+0.14) but a lower precision (-0.14). Relevant generated keywords from each LLMs are different (and so combining them allows to increase the recall) whereas the irrelevant generating keywords are more frequently the same ones leading

so to a lower degradation in the precision (*e.g.* increase of false positives is lower than the increase of true positives when aggregating the keywords). 4938 more text extracts are correctly mapped to their associated SDGs using ULA.

Figure 6 compares the AA and ULA approaches in terms of precision and recall per SDG. As expected from previous results, the improvement in recall comes with a degradation of the precision. As the worst example with SDG15, the precision is lowered by 26% with a limited increase of 15% for the recall. Globally, the improvement in recall is higher that the degradation of the precision

On average, ULA correctly classifies 309 more text extracts per SDG. We have also observed that performance may fluctuate depending on the SDG. ULA offers better performance for SDG3, with only 7.56% of additional text extracts classified in addition to those classified correctly by AA. In contrast, results for SDG9 show a higher recall of 19.13%. However, regarding other approaches, we observed that AA does better than any individual LLM for 4 SDGs out of 16. Considering only LLMs, PaLM-based keyword generation is the most accurate for 15 SDGs out of 16. This is all the more interesting since PaLM is the LLM that generated the fewest new keywords among the 3 LLMs.

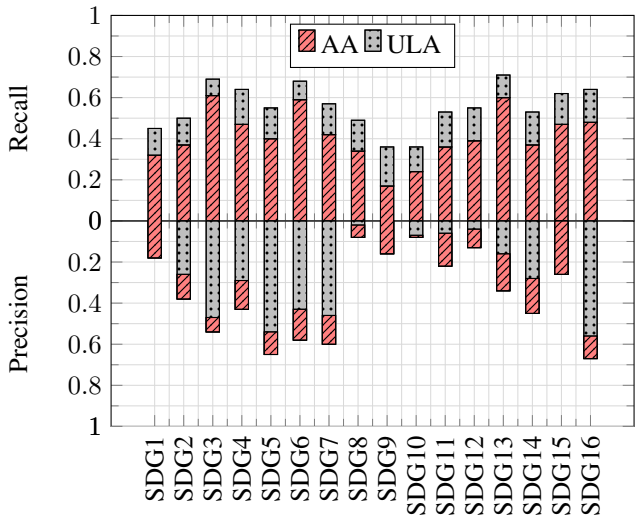

Fig. 6. Precision and recall performances (AA vs ULA)

VI. Conclusion

In this paper, we propose a method to associate texts with SDGs using LLMs. Our results highlight the complementary nature of queries containing keywords manually validated and reviewed by experts, on the one hand, and queries containing keywords automatically generated by LLMs, on the other hand. The large learning sets used to build LLMs ensure linguistic diversity.

We have also shown that the overall performance (F1-score) of queries built and validated manually by experts is globally comparable to that obtained using an extended approach such as ULA based on the aggregation of AA and 3 LLM outputs.

However ULA allows to create a set of keywords with 7 times larger than AA alone. Although it is able to retrieve more relevant text extracts which was our initial goal, it leads to degrade the precision. In practice, a good trade-off must be carefully set based on the ultimate objectives of the analysis.

We have also observed that the fine-tuned versions of LLMs dedicated to conversational tasks contain safety filters that prevent them from responding to certain requests that are deemed unethical or inappropriate (*e.g.* when generating synonyms for the keyword *Child Abuse*). However, these filters can be explicitly deactivated for our future work.

There are also additional avenues for improvement to be explored. The ULA set can be composed only of keywords common to the different LLMs, after implementing a voting mechanism (rather than blind aggregation).

Acknowledgment

We would like to thank the experts in sustainable finance of our partner, Banque de Luxembourg S.A. for their involvement and advice in carrying out this research.

References

[1] United Nations, "Transforming our world: the 2030 agenda for sustainable development," 2015. [Online]. Available: https://sdgs.un.org/2030agenda

[2] Y. Kashnitsky, G. Roberge, J. Mu, K. Kang, W. Wang *et al.*, "Evaluating approaches to identifying research supporting the united nations sustainable development goals," *arXiv*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2209.07285

[3] W. Wang, W. Kang, and J. Mu, "Mapping research to the sustainable development goals: A contextualised approach," *Research Square*, 2023. [Online]. Available: https://doi.org/10.21203/rs.3.rs-2544385/v3

[4] V. Maurice, O. René, and S. Eike, "Search Queries for "Mapping Research Output to the Sustainable Development Goals (SDGs)" v1.0," *Zenodo*, Jan. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.3817352

[5] ——, "Search Queries for "Mapping Research Output to the Sustainable Development Goals (SDGs)" v5.0.2," *Zenodo*, Jul. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4883250

[6] V. Maurice, S. Eike, and G. Yassin, "Survey data of "Mapping Research Output to the Sustainable Development Goals (SDGs)"," *Zenodo*, May 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3813230

[7] J. Bamini, B. Roy, A. Kevin, and K. Lisette, "Identifying research supporting the united nations sustainable development goals," *Elsevier Data Repository, V1*, 2019. [Online]. Available: https://doi.org/10.17632/87txkw7khs.1

[8] R. Maxime, K. Yury, B.-V. Alexandre, C. David, K. Paul *et al.*, "Improving the scopus and aurora queries to identify research that supports the united nations sustainable development goals (sdgs)," *Elsevier Data Repository, V4*, 2021. [Online]. Available: https://doi.org/10.17632/9sxdykm8s4.4

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[10] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao *et al.*, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," *arXiv*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.10420

[11] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.09288

[12] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin *et al.*, "Palm 2 technical report," *arXiv*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.10403

[13] OSDG, UNDP IICPSD SDG AI Laband PPMI, "Osdg community dataset (osdg-cd)," *Zenodo*, Jul. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8107038