BLU-SynTra: Distinguish Synergies and Trade-offs Between Sustainable Development Goals Using Small Language Models

Loris Bergeron, Jérôme François, Radu State, Jean Hilger
Banque de Luxembourg, University of Luxembourg - SnT SEDAN & Finnovation Hub
14 Bd. Royal L-2449 Luxembourg, 29 Av. John F. Kennedy L-1855 Luxembourg
loris.bergeron@bdl.lu, {jerome.francois, radu.state, jean.hilger}@uni.lu

Abstract

Since the United Nations defined the Sustainable Development Goals, studies have shown that these goals are interlinked in different ways. The concept of SDG interlinkages refers to the complex network of interactions existing within and between the SDGs themselves. These interactions are referred to as synergies and trade-offs. Synergies represent positive interactions where the progress of one SDG contributes positively to the progress of another. On the other hand, trade-offs are negative interactions where the progress of one SDG has a negative impact on another. However, evaluating such interlinkages is a complex task, not only because of the multidimensional nature of SDGs, but also because it is highly exposed to personal interpretation bias and technical limitations. Recent studies are mainly based on expert judgements, literature reviews, sentiment or data analysis. To remedy these limitations we propose the use of Small Language Models in addition of an advanced Retrieval Augmented Generation to distinguish synergies and trade-offs between SDGs. In order to validate our results, we have drawn on the study carried out by the European Commission's Joint Research Centre which provides a database of interlinkages labelled according to the presence of synergies or trade-offs.

Keywords: United Nations (UN), Sustainable Development Goals (SDGs), Small Language Models (SLMs), Retrieval Augmented Generation (RAG), Mistral, Orca 2, Phi-2, Generative Query Reformulation (GenQR), Context Aware Query Rewriting (CAR), Reciprocal Rank Fusion (RRF), Zero-Shot Classification

1. Introduction

In 2015, the agenda dedicated to sustainable development was adopted by all 193 member states of the United Nations (UN)(United Nations and Development, 2015). A set of 17 Sustainable Development Goals (SDGs) was defined and reported in Table 1. The establishment of these 17 SDGs, broken down into 169 targets and 232 indicators, would have us isolate all these elements as if, in theory, no interlinkages were possible between the economic, social and governance dimensions. As an example, assuming *SDG3 Good Health and Well-being* and *SDG12 Responsible Consumption and Production*, there is no clear assessment if these SDGs present synergies or trade-offs. In other words, would having a positive impact on *SDG3* also mean having a positive impact on *SDG12* and vice versa? At a first glance, having a positive impact on *SDG12* seems to have a positive impact on the health and well-being of populations. However, if we improve *SDG12* on responsible consumption and production, carbon footprint can go down also. In that case, this might lead to a trade-off with *SDG13 Climate Action* and with *SDG7 Affordable and Clean Energy*. Obviously, this reasoning is based on personal beliefs that are unique to each individual and is therefore, by definition, subject to personal bias. The complexity of these interlinkages is all the more true if we opt for a finer granularity by appealing to the SDGs targets. In this case we have a combination of potential 14196 interlinkages. It is all the more essential to obtain an overview of these interlinkages to give policy-makers all the transparency to make the right decisions to successfully implement these objectives. Understanding the range of positive and negative interlinkages among the SDGs is the key to unlocking their full potential while ensuring that progress in some dimensions does not have a negative impact on others(noa, 2017). Hence, this paper introduces a method capable to automatically distinguish synergies and trade-offs in the interlinkages of SDGs using Small Language Models (SLMs) thanks to their cognitive capacities. In particular, we are interested in reproducing results established by experts in scope of a research(European Commission. Joint Research Centre., 2023) which is part of *KnowSDGs*[1] and carried out by the European Commission's Joint Research Centre (JRC). The database provided in this study brings together a number of interlinkages at goals and targets levels. For many months now, the research on Large Language Models (LLMs) has continued to progress. Transformers architecture(Vaswani et al., 2017) were considered to be LLMs regardless of the number of training parameters included in them. We used them mainly for their cognitive capacities but also and above all for their vast knowledge since they were trained on

[1] https://knowsdgs.jrc.ec.europa.eu

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 21–33

impressive volumes of data. However, since the research carried out by Microsoft(Eldan and Li, 2023), a distinction can be made between LLMs and SLMs. We can therefore consider as an SLM an LLM with a far smaller number of parameters, several hundred billion against a few billion. SLMs are not used for their knowledge, but rather for their impressive cognitive capacities given their small size. Recent advances in Generative AI (GenAI) have opened up new possibilities in the field of SLMs which are now used in a multitude types of tasks. In this paper, we promote the use of SLMs to replicate the results obtained in JRC's study. The obtained results shows their ability to distinguish synergies and trade-offs between SDGs targets. This type of usage can be industrialised, but is also made close through with the help of a relevant context, since such an analysis must be carried out given a specific environment (e.g. political, economic, geographical)(Le Blanc, 2015). Our contribution to scientific research in relation to these SDG themes can be broken down into four areas:

- An innovative methodology, based on the use of SLMs and an advanced RAG (Retrieval Augmented Generation) (Lewis et al., 2021) workflow, to distinguish synergies and trade-offs between SDGs targets in a set of documents
- An open architecture that can be replicated by research teams or companies while still having access to infrastructure with limited computing and hardware power and hosted internally for governance reasons
- An implementation of the aforementioned architecture using Mistral 7b (Mistral)(Jiang et al., 2023), Orca 2 7b (Orca)(Mitra et al., 2023), Phi-2 2.7b (Phi)(Javaheripi et al., 2023)
- A comparative analysis of our results based on the study carried out by the European Commission(European Commission. Joint Research Centre., 2023)

The structure of the paper is organized as follows: Section 2 provides an overview of related work. Our method, called BLU-SynTra is described in Section 3. Then, details of the validation set used to confirm our results are presented in Section 4, comparative analysis and the results we achieved are presented in Section 5. Lastly, Section 6 provides concluding remarks on the conducted research and suggests potential enhancements for future research.

2. Related Work

In 2015, when SDGs were conceptualized by the UN, research topics related to the identification of connections between the SDGs began to appear(Le Blanc, 2015). The identification of connections between SDG targets is carried out on the basis of a manual semantic analysis by determining that if two targets refer to the same global concept, they can be assumed to be interlinked. Obviously, this method is highly exposed to fluctuations in human interpretation. In 2017, the International Council for Science (ICSU)(noa, 2017) published a report to explore the nature of interlinkages between SDGs. The evaluation method is based on assigning manually a score to quantify the interlinkages. Therefore, this evaluation is based on expert opinion and a review of the literature. At European level, in 2019, the European Commission's Joint Research Centre (JRC) published a first version of a research(European Commission. Joint Research Centre., 2019) highlighting interlinkages in order to ensure policy coherence in relation to the SDGs, based on a literature review. Hereafter, more and more related research has been carried out (Bali Swain and Ranganathan, 2021; Fariña García et al., 2021; Dawes, 2022; Dawes et al., 2022; Song and Jang, 2023). Use of new methods, like analysis methods based on correlation networks or semantic analysis networks, to determine interlinkages between SDGs are being used. These research does not attempt to distinguish, from a qualitative point of view, the interlinkages type when they are actually present. In 2023, the JRC published a new research(European Commission. Joint Research Centre., 2023) to review the progress of work on the existence of synergies or trade-offs in interlinkages between SDGs in different contexts. Based on this work, a database of interlinkages is established through a literature review. This database provides the community with

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 Prod.
SDG13	Climate Action
SDG14	Life Below Water
SDG15	Life on Land
SDG16	Peace, Justice and Strong Institutions
SDG17	Partnerships for the Goals

Table 1: The 17 Sustainable Development Goals

a list of 18780 interlinkages, each qualified as a synergy or a trade-off alongside the method used to assert it. As highlighted previously, past work mainly relies on experts judgments, literature review or data analysis methods to explore SDG interlinkages. As the current state of the art(Issa et al., 2024) does not refer to a methodology based on SLMs to qualitatively distinguish the type of interlinkages, our research aims to explore the potential benefit of such models.

3. Method

3.1. BLU-SynTra overview

The overall process of BLU-SynTra consists of adaptation and combination of different methods and practices as represented in Figure 1. They are also detailed in the following sub-sections. The first building block in the figure is *Optimised data indexing*. This block takes as input a set of documents related to various studies or reports, where interlinkages (i.e. synergies or trade-offs) between SDGs are explained and validated by experts. These documents are then handled by an unstructured data ingestion mechanism to extract the information they contain. Then, a series of processing steps create chunks, using a parent-child strategy and static thresholds to divide up the information. These chunks are then summarised using a SLM (Mistral, Orca and Phi) to retrieve their meaning by reducing their context size. Chunks were later incorporated into a vector database using the best performing model at the time of our research to perform Semantic and Textual Similarity (STS) operations (on the basis of information established by the Massive Text Embedding Benchmark (MTEB)[2]).

The next building block is *Information Retrieval* whose aim is to contextualize a user query given as input about interlinkages based on the previously indexed document using RAG (Retrieval Augmented Generation). Advanced RAG methods(Gao et al., 2024) like Generative Query Reformulation (GenQR)(Wang et al., 2023b), Context Aware Query Rewriting (CAR)(Anand et al., 2023), Reciprocal Rank Fusion (RRF)(Cormack et al., 2009) have been used to maximise the final results of our research. Finally, *Distinguish interlinkages* relies on the extracted context to apply common sense reasoning and understanding of language capabilities offered by SLMs to distinguish synergies and trade-offs in interlinkages available in our validation set using Zero-Shot (ZS)(Brown et al., 2020) classification.

[2]https://hf.co/spaces/mteb/leaderboard

3.2. Optimised data indexing

3.2.1. Ingest unstructured data

When processing unstructured data, the challenge is to extract the contained elements as faithfully as possible to avoid any analysis errors.

Let's define D as the set of documents used in our research:

$$D = \{d_1, d_2, \ldots, d_n\}$$

Within each d_i, we assume the elements e:

$$d_i = \{e_1^i, e_2^i, \ldots, e_{m_i}^i\}$$

Where i is the index of the document within the set D and m_i is the number of textual elements in document d_i.

To achieve this, we used the well known *unstructured* [3] library that includes an OCR model to segment a document and extract its content. At this stage of the process, our aim was not to extract any information that has already been chunked or organised, but only to extract $e_{m_i}^i$ as represented in the original d_i document excluding images and tables. This results in a set of elements of different element types *(e.g. title, page_break, footer, etc.)*. To focus only on information having semantic value, only *NarrativeText* typed elements are kept. They consist of text composed of at least two sentences. Assuming $narrative(d_i) \subset d_i$ is only the remaining *NarrativeText* elements of d_i, we refine D as D':

$$D' = \{narrative(d_i)\}, d_i \in D$$

3.2.2. Chunking elements in parent-child

Once narrative text is extracted, it is essential for our solution to conserve the related context and meanings. We have decomposed each element e_j^i using a parent-child strategy in which the elements can be made up of several parents p and several smaller children c implemented in the *RecursiveCharacterTextSplitter* function from *LangChain*[4]. This text splitter is suggested for general text and it uses a list of default separators (i.e. \n\n, \n, space, char), aiming to maintain paragraphs, then sentences, and finally words together as much as possible since they are viewed as the most semantically connected elements. The maximal *chunk_size* parameters for parents and children have been set to 4096 and 2048 respectively based on the maximum window context size of the SLMs we used. We can therefore establish that each element e_j^i can be represent as the set of children: $child_{e_j^i} = \bigcup_{p \in P_{i,j}} \{p_k\}$ with $P_{i,j}$ is the set

[3]https://github.com/unstructured-io/unstructured
[4]https://python.langchain.com

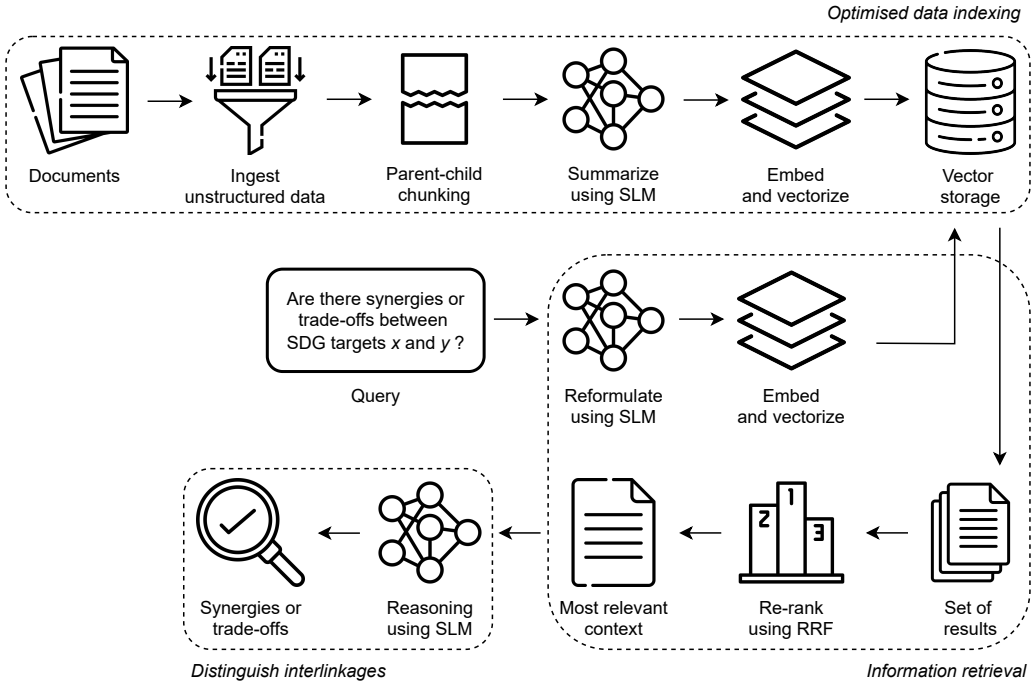

Figure 1: BLU-SynTra overview

of parents derived from e^i_j and p_k is the *k*-th child of the parent p.

As a result, we define the whole set of children to represent the original set of documents D as:

$$C = \bigcup_{d \in D', e^i_j \in d} child_{e^i_j}$$

3.2.3. Summarize chunks

Advanced RAG methods promotes the principle of summarization to improve the ability of LLMs to understand key information, particularly when dealing with extensive texts(Gao et al., 2024). BLU-SynTra thus includes such processing as well.

Creating informative summaries based on a longer text is a quite complex, unlike summarising smaller texts, which justifies our previous breakdown in section 3.2.2.

Each $c \in C$ is summarized using the different SLMs. sm^a_c is the generated summary for c using the SLM $a \in SLM$ with $SLM = \{Mistral, Orca, Phi\}$ resulting in the full summarization of all documents:

$$Sm_c = \bigcup_{c \in C} sm^a_s(c), a \in \{SLM\}$$

At the end of the generation process we have a set of 15009 summaries (5003 for each SLM).

The SLMs are conditioned to produce summaries as faithful and consistent as possible with our various $c \in C$ in order to minimise errors in the following way:

> *Please provide a summary of the following text. Ensure the summary is clear, coherent, and faithful to the content of the original text.*
> *Text:* < c >

For a clearer understanding of how this summary stage works, we have appended an example A of a randomly selected child c and the corresponding summary Sm_c produced. To remain as neutral as possible, we did not modify the parameters within the SLMs *(e.g. temperature, top_p)* and used the same prompt for each of them.

3.2.4. Embeddings creation and storage

Once summaries are created, they are stored as embeddings to enable easy comparison between them. BLU-SynTra relies on the Universal AnglE Embedding(Li and Li, 2023) in Large-V1 version(UAE-Large-V1)[5] as the embedding model. At the time of this research, this is the most advanced model to perform STS operations in English. Given the technical specificities of the model, chunks are embedded in 1024-long vector. All vectors are stored in chroma[6]. Default use of the Hierarchical Navigable Small World (HNSW)(Malkov and Yashunin, 2020) method in chroma, coupled with the use of the cosine function to perform similarity operations allows us to retrieve the appropriate information. In chroma's *documents* field, we have

[5] https://hf.co/whereisai/uae-large-v1
[6] https://docs.trychroma.com

stored all the Sm_c summaries along side the child c used to create them, their relative parent p and the source document d_i as metadata. Each Sm_c is so associated with a vector representation noted v_{Sm_c}:

$$\mathbf{v}_{Sm_c} = \begin{pmatrix} v_{Sm_c1} \\ v_{Sm_c2} \\ \vdots \\ v_{Sm_c1023} \\ v_{Sm_c1024} \end{pmatrix}$$

3.3. Information retrieval

3.3.1. Query reformulation

To retrieve information, we generate a query. Assuming an initial query q, the objective is to determine if there are synergies or trade-offs between SDG targets, as for instance:

> *q: Are there synergies or trade-offs between SDG targets 17.11 and 10.7?*

As the formulation of a query to a generative AI model can have a significant impact on the final classification result, the principle of reformulation is widespread in Information Retrieval (IR) problems and is used to counter problems linked to a more or less extensive vocabulary. To optimise our results, we used existing reformulation mechanisms(Wang et al., 2023b; Anand et al., 2023). On the one hand, a reformulation noted as $GenQ$ is solely based on the cognitive capacities of SLMs. On the other hand we also define $GenQCAR$ as a reformulation based on particular context related to the SDG targets helping SLMs in their reformulation task. While $GenQ$ reformulation simply reformulates and expands q, the $GenQCAR$ approach enriches knowledge by providing it with the definitions of synergy and trade-off as defined in the JRC study as well as the definitions of the two targets as defined by the UN[7]. In the case of $GenQCAR$, the additional information made available to the SLMs is transmitted to it when q is reformulated using a prompt B specifically written for this purpose. In the appendix C, two examples of $GenQ$ and $GenQCAR$ are given and have been derived using q mentioned earlier. A higher vocabulary richness can be observed in the case of $GenQCAR$, but also and above all the use and understanding of the terms synergy and trade-off in accordance with the definitions given by the JRC.

P denotes the reformulation process and can therefore define for each q the process:

$$P(q) = \{GenQ(q), GenQCAR(q)\}$$

[7] https://unstats.un.org/sdgs/metadata

During the IR step, we obtained a set of results for which we retrieve the 10 most similar items by query, defined as follows:

$$R_{total}(q) = R_q \cup R_{GenQ(q)} \cup R_{GenQCAR(q)}$$

3.3.2. Re-rank result sets

When retrieving information from R_{total}, the result is a set of elements associated with cosine similarity scores in R_i. In order to identify the most recurrent documents in R_{total}, we used the Reciprocal Rank Fusion (RRF)(Cormack et al., 2009) method. In contrast to individual ranking methods, the authors have shown that the RRF method is capable of consistently obtaining better results than the standard Condorcet Fuse method(Montague and Aslam, 2002). RRF weights each document in R_i with the inverse of its position on the rank. It thus gives preference to documents at the *top* of the rank and penalizes documents below the *top* of the rank. In addition, this approach is unsupervised, that is also a significant advantage to be applicable. RRF therefore sort our set R_{total} according to a scoring formula based on a set of rankings R_i:

$$\text{RRFscore}(r \in R_{total}) = \sum_{r \in R_i} \frac{1}{k + r(d)} \quad (1)$$

with $r(d)$ the rank of document d and k a parameter, set to $k = 60$ as suggested in the original paper of RRF (Cormack et al., 2009).

Finally, the document with the highest score is considered to be the most appropriate given the queries formulated in the previous step. Thanks to what we have seen in section 3.2.4, this enables us to identify the associated p and refer to as the *most relevant context* to be used by the SLM to carry out the classification step.

3.4. Distinguish interlinkages

To be able to distinguish the interlinkages type between the two targets concerned, we decided to use a Zero-Shot (ZS)(Brown et al., 2020) classification. We used ZS by augmenting the knowledge of the model with the definitions of *synergy* and *trade-off* as defined in the JRC study and also with the *most relevant context* retrieved in the previous step. Thanks to context augmentation, the knowledge of SLMs is increases and enables performing reasoning tasks and thus determine, in the given context, the type of interlinkages present between the two targets. The fact that we add in our prompt some more detailed background information (i.e. most relevant context) as well as the definitions of the two classes to be classified (i.e. synergy and trade-off) improves the accuracy of the predictions(Wang et al., 2023c). We have define the

prompt detailed in appendix D to carry out this operation. In addition to the classification, we request a justified explanation for the underlying reasoning behind it. Such kind of explanation could be made available to a decision-maker to obtain all the transparency needed to understand these interlinkages. To illustrate this process, an example of output is given in appendix E resulting from the classification between targets 6.a and 10.b of our validation set.

We could have implemented methods like Zero-Shot Chain of Thought (ZS-CoT)(Kojima et al., 2023) or Clue And Reasoning Prompting (CARP)(Sun et al., 2023). These methods, used to classify texts using LLMs, compensates for the models' lack of reasoning capacity by adopting a progressive reasoning strategy to overcome these limitations in complex environments. However, the related literature highlights that the added value of such methods, based on progressive reasoning, is correlated with the size of the model used. In our case, by the limited size of the number of parameters in our SLMs, the added value in terms of reasoning is not significant and would negatively increase classification processing time.

4. Experimental setup

4.1. Selected SLMs

We chose three SLMs which considered to be the most common from the state-of-the-art at the time of our research. We also selected multiple models for comparative analysis rather than pre-selecting one. However, they are used independently and can be interchanged. In other words, BLU-SynTra used in production would rely on the use of the SLM that exhibits the most efficient summarisation behavior, as discussed in section5:

- Mistral 7b(Jiang et al., 2023) - Designed to use Grouped-Query Attention (GQA)(Ainslie et al., 2023) and Sliding Window Attention (SWA)(Beltagy et al., 2020)(Child et al., 2019). The use of GQA and SWA allows us to significantly accelerate the inference speed while reducing the memory required for the decoding phase. This choice is particularly well suited to infrastructures with limited computing power.

- Orca 2 7b(Mitra et al., 2023) - Based on the architecture of Llama-2(Touvron et al., 2023). This version 2 of Orca has the advantage of employing a varied number of reasoning techniques (e.g. step-by-step, recall then generate, recall-reason-generate, etc.) while being able to choose the right method for a given task.

- Phi-2 2.7b(Javaheripi et al., 2023) - Builds on the work of the previous version, Phi1.5(Li et al., 2023). This version currently shows similar or better cognitive performance than models with 13b parameters or less. While its parameter size is more than half that of the two previous SLMs, its main innovation lies in the use of *textbook-quality* data(Gunasekar et al., 2023) and the addition of new synthetic data. This new version uses an innovative method of knowledge transfer to accelerate its training speed while delivering superior benchmark scores compared to the previous version.

4.2. Validation set

As stated earlier, we rely on the database provided by the JRC serves. Since this database is the result of work carried out by multiple JRC experts to avoid individual bias. We thus consider this database enough accurate to serve as a validation set. At the SDG target level, there is a total of 10614 interlinkages but only 5715 are unique. There are 80.5% synergies, 10% trade-offs and 9.5% not specified resulting in a significant imbalance between classes. For the remainder of our research, only interlinkages specifically associated *synergy* or *trade-off* type are kept. In addition, we excluded interlinkages whose *clear_direction* variable was set to *no*. By applying these quality filters we obtain a set of 4682 interlinkages, of which 2956 are unique, divided into 4172 (89.1%) synergies and 510 (10.9%) trade-offs. In order to optimise our experiment, we randomly sampled this group to keep only 10% of the total. This brings our total number of classes to 468, divided into 419 (89.53%) synergy classes and 49 (10.47%) trade-off classes. Regarding to the methods of analysis used to establish the distinctions between synergy and trade-off in the database, no filter has been applied resulting in the breakdown shown in Table 2. To compare and replicate our results, we have made our final validation set available online[8].

We looked at the distribution of classes according to the targets selected in our validation set. For sake of clarity, targets are grouped by SDG they relate to. In Figure 2, the distribution of synergies and trade-offs is presented.

5. Results

5.0.1. Evaluation of summaries

This first experiment aims at assessing the quality of the summarization process which is critical for the IR process. ROUGE(Lin, 2004) metric might have been used to evaluate the quality of generated summaries. This metric measures the similarity between a summary sm_c^a in comparison to the refer-

[8]https://github.com/lrsbrgrn/blu-syntra

N	Method of analysis	Synergy	Trade-off
1	Data Analysis	8	2
2	Expert judgement	133	9
3	Literature review	48	4
4	Mixed (Expert judgement & Data analysis)	8	1
5	Mixed (Literature review & Data analysis)	25	2
6	Mixed (Literature review, Expert judgement & Data analysis)	3	0
7	Mixed (Literature review, Expert judgement & Modelling)	1	4
8	Mixed (Literature review & Expert judgement)	174	21
9	Mixed (Semantic analysis, Literature review & Expert judgement)	15	6
10	not_specified	4	0
Total		419	49

Table 2: Distribution of classes by analysis method

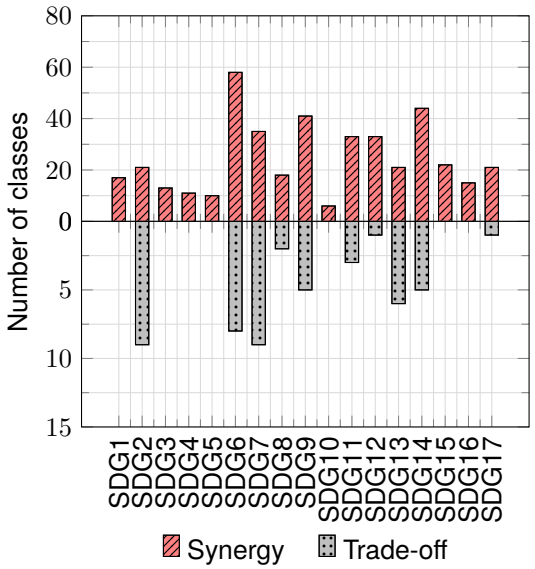

Figure 2: Distribution of classes by SDG

ence noted as c. We could have carried out our evaluation using ROUGE-N (1 and 2) to measure the proportion of common unigrams and bi-grams. In addition, ROUGE-L metric would have enabled us to evaluate the longest sequence of words shared between our summary and its reference. A long shared sequence indicates a definite similarity between the two. However, metrics such as ROUGE, although widely used in NLP tasks, show low correlations with human judgements(Wang et al., 2023a). Based on this assessment of the state of the art we used the G-Eval(Liu et al., 2023) framework. The latter relies on an LLM, in our case GPT-4(OpenAI et al., 2023), as an evaluator to determine several metrics (i.e. relevance, coherence, consistency and fluency). Table 3 shows the different results obtained by G-Eval. We can observe very similar results between Mistral and Orca but also excellent results for Phi given its very small size. However, Mistral is superior to Orca in 3 out of the 4 metrics, and to Phi in all cases. With Mistral, we observed

Metrics	Mistral	Orca	Phi
Relevance	**4.6**	**4.6**	3.9
Coherence	**4.5**	4.4	4.0
Consistency	**4.9**	4.8	4.1
Fluency	**3.0**	2.9	2.8

Table 3: Evaluation using G-Eval

an average length of 222 words for each c compared with an average length of 83 words for the summaries produced. This is equivalent to dividing the size of the text by almost 3 and thus justifies the use of an advanced RAG method to reduce the text to retain only the key information.

5.0.2. Validity of classifications

This experiment evaluates to which extent our approach can automatically infer if synergies or trade-offs exist between SDG goals. For this analysis phase, only *Mistral* is used due to its highest scores on summarization as evaluated in the previous section. As a first experiment, we were interested in assessing the validity of our classifications and the underlying behaviour of the SLM according to a binary classification where the positive label $pos\_label = SYNERGY$. The results show a very good capability of BLU-SynTra to identify the synergies with $F1\_score = 0.88$, $Precision = 0.92$ and $Recall = 0.84$. However, a deeper look at the confusion matrix in the Table 4 highlights a bias in overestimating these synergies and, as a an opposite effect, a notable difficulty in identifying trade-offs. Of the 49 trade-offs available in our validation set, only 20 (40.82%) were actually correctly identified. Our validation set shows a strong asymmetry in the classes it contains, since synergy and trade-off represent 89.53% and 10.47% of the whole respectively. This result still highlights the difficulties encountered by Mistral in producing classifications for which the finesse of the language, the subtlety of the words and the intonations present challenges to their reasoning function.

		Predicted	
		Synergy	Trade-off
Actual	Synergy	354	65
	Trade-off	29	20

Table 4: Confusion matrix

Secondly, results were differentiated according to the SDG each target they relate to. We found significantly heterogeneous performance metrics for SDGs 2, 7 and 13, with F1 scores of $F1_{SDG2} = 0.70$, $F1_{SDG7} = 0.89$ and $F1_{SDG13} = 0.79$ respectively. SDGs 2 and 7, as shown in Figure 2, are among the largest contributors to trade-offs. For SDG 2, only 3 of the 9 trade-offs in our validation set were correctly classified as such. Regarding SDG 7, only 2 of the 9. Notably, SDG 6 is the third highest contributor of trade-offs in our validation set but still has an $F1_{SDG6} = 0.92$ with 5 of the 8 trade-offs correctly identified. This generally highlights a high divergence of BLU-SynTra capabilities among SDG.

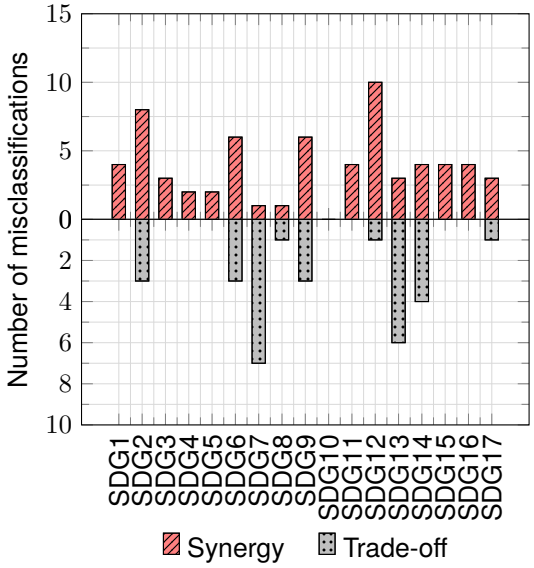

Figure 3: Misclassification by SDG

The last experiment further investigate results obtained according to the analysis methods used by the JRC to determine the presence of synergy or trade-off (see Table 2). As can be observed in Table 5, there is no consistency between the results obtained. In general, a significant deterioration in F1 scores for the *Mixed (Expert Judgement & Data analysis)* (M-EJDA) and *Data Analysis* (DA) methods can be observed. These two approaches lead to the worst performance. However, although the *Expert Judgement* (EJ) analysis method is the second method with the most number of classes, the F1 score obtained is the highest. We have also observed that the *Literature Review* (LR) method is well approximated by BLU-SynTra with all the trade-offs correctly identified.

Methods	**Precision**	**Recall**	**F1-score**
M-EJDA	0.80	0.50	0.62
DA	0.71	0.63	0.67
EJ	0.97	**0.92**	**0.94**
LR	**1.00**	0.73	0.84

Table 5: Metrics by analysis method

This raises questions about mixed approaches compared with single approaches (i.e. using only one analysis method). We have noted *Mixed* the analysis methods employing several sub-methods, and noted *Single* the methods employing only one analysis method. In Table 2, the *Mixed* methods are identified by the prefix *(Mixed)*, the others are consequently attached to the *Single* category. We therefore observed a slight superiority when comparing *Single* and *Mixed* approaches (see Table 6). However, the *Single* approaches were able to correctly identify 60.00% of the trade-offs, unlike the *Mixed* approaches, which were only able to obtain a score of 32.35% and therefore leads to a deterioration at the global level of the classification metrics.

Methods	**Precision**	**Recall**	**F1-score**
Single	**0.97**	**0.86**	**0.91**
Mixed	0.89	0.83	0.86

Table 6: *Single* and *Mixed* metrics

6. Conclusion

In this paper, we have proposed a complete solution entitled BLU-SynTra relying on SLMs to identify synergies and trade-offs between SDG targets. We have shown that traditional ZS text classification methods, enhanced by a context and definitions retrieved using several advanced RAG concepts, can make it easy to identify synergies and justify to decision-makers the underlying reasoning behind this distinction in a given environment. However, the identification of trade-offs lacks precision, and most of all with high variability according to the considered analysis method or SDG. Linguistic complexity and subtle vocabulary make it difficult for SLMs to identify trade-offs and distinguish them from synergies. Despite this, this first research work aims to open up new possibilities for using SLMs to carry out this interlinkages classification task as we have experienced and, more generally, in tasks requiring complex reasoning to be carried out in infrastructures with limited hardware resources or at lower cost that LLMs. New advances in summary generation (Zhang et al., 2023)

will be the subject of future improvement to create summaries in an iterative way in order to reduce as much as possible the errors and hallucinations induced by SLMs. In terms of reasoning skills, our plan is to leverage ReAct(Yao et al., 2023) in order to compare the results obtained with those obtained in this research. ReAct seem to indicate better performance than standard approaches, even for models with very small parameter sizes.

7. Acknowledgment

We would like to thank the sustainable finance experts at Banque de Luxembourg from the Private Banking Investments team, as well as the sustainable finance experts at Banque de Luxembourg Investments for their involvement and advice in carrying out this research. We would also like to thank the teams at the Joint Research Centre (JRC) who made themselves available to answer our questions when we needed them most.

8. Bibliographical References

2017. A guide to SDG interactions: from science to implementation. Technical report, International Council for Science (ICSU).

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. ArXiv:2305.13245 [cs].

Abhijit Anand, Venktesh V, Vinay Setty, and Avishek Anand. 2023. Context Aware Query Rewriting for Text Rankers using LLM. ArXiv:2308.16753 [cs].

Ranjula Bali Swain and Shyam Ranganathan. 2021. Modeling interlinkages between sustainable development goals using network analysis. *World Development*, 138:105136. .

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv:2004.05150 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. ArXiv:1904.10509 [cs, stat].

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759, Boston MA USA. ACM.

J.H.P. Dawes. 2022. SDG interlinkage networks: Analysis, robustness, sensitivities, and hierarchies. *World Development*, 149:105693. .

Jonathan H. P. Dawes, Xin Zhou, and Mustafa Moinuddin. 2022. System-level consequences of synergies and trade-offs between SDGs: quantitative analysis of interlinkage networks at country level. *Sustainability Science*, 17(4):1435–1457. .

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? ArXiv:2305.07759 [cs].

European Commission. Joint Research Centre. 2019. *Interlinkages and policy coherence for the sustainable development goals implementation: an operational method to identify trade offs and co benefits in a systemic way.* Publications Office, LU.

European Commission. Joint Research Centre. 2023. *Uncovering SDG Interlinkages: interconnection at the core of the 2030 Agenda : an analysis of the state of the art on SDG Interlinkages and an update of the JRC tool to foster policy coherence for sustainable development in EU policymaking.* Publications Office, LU. .

María Consuelo Fariña García, Víctor Luis De Nicolás De Nicolás, José Luis Yagüe Blanco, and Jesús Labrador Fernández. 2021. Semantic network analysis of sustainable development goals to quantitatively measure their interactions. *Environmental Development*, 37:100589. .

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv:2312.10997 [cs].

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. ArXiv:2306.11644 [cs].

Lea Issa, Toufic Mezher, and Mutasem El Fadel. 2024. Can network analysis ascertain SDGs interlinkages towards evidence-based policy planning? A systematic critical assessment. *Environmental Impact Assessment Review*, 104:107295.

Mojan Javaheripi, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, and Caio Mendes. 2023. The Surprising Power of Small Language Models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. .

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. ArXiv:2205.11916 [cs].

David Le Blanc. 2015. Towards Integration at Last? The Sustainable Development Goals as a Network of Targets. *Sustainable Development*, 23(3):176–187.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv:2005.11401 [cs].

Xianming Li and Jing Li. 2023. AnglE-optimized Text Embeddings. ArXiv:2309.12871 [cs].

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks Are All You Need II: phi-1.5 technical report. ArXiv:2309.05463 [cs].

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. ArXiv:2303.16634 [cs].

Yu A. Malkov and D. A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: a Survey. ArXiv:2302.07842 [cs].

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching Small Language Models How to Reason. ArXiv:2311.11045 [cs].

Mark Montague and Javed A. Aslam. 2002. Condorcet fusion for improved retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 538–548, McLean Virginia USA. ACM.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua

Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. ArXiv:2312.05934 [cs].

Jaemin Song and Chang-Ho Jang. 2023. Unpacking the sustainable development goals (SDGs) interlinkages: A semantic network analysis of the SDGs targets. *Sustainable Development*, 31(4):2784–2796.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text Classification via Large Language Models. ArXiv:2305.08377 [cs].

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Shreya Pathak, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher Choquette, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clement Farabet,

Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma. Publisher: Kaggle.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem 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, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng 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:2307.09288 [cs].

Department of Economic United Nations and Social Affairs-Sustainable Development. 2015. Transforming our world: the 2030 Agenda for Sustainable Development. ISBN: A/RES/70/1 Pages: 16301 Type: General Assembly.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. ArXiv:2303.04048 [cs].

Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2023b. Generative Query Reformulation for Effective Adhoc Search. ArXiv:2308.00415 [cs].

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023c. Large Language Models Are Zero-Shot Text Classifiers. ArXiv:2312.01044 [cs].

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Fine-tuned Language Models Are Zero-Shot Learners. ArXiv:2109.01652 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. ArXiv:2210.03629 [cs].

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. SummIt: Iterative Text Summarization via ChatGPT. ArXiv:2305.14835 [cs].

A. Example of a generated summary

c: in terms of meeting most of the SDGs by 2030. Fast population growth slow progress toward sustainable production and consumption and the negative impacts of the ongoing coronavirus disease 2019 COVID19 pandemic on socioeconomic development make it difficult for China to meet the SDGs. Therefore effective water pollution control in China is urgently needed. Here we analyze the interactions

Sm_c: China faces challenges in achieving most of the Sustainable Development Goals (SDGs) by 2030 due to a fast-growing population, slow progress towards sustainable production and consumption, and the negative impacts of COVID-19 on socio-economic development. Amidst these challenges, effective water pollution control in China is particularly urgent. This analysis explores the interactions between these issues.

B. Prompt to reformulate queries

$GenQ$: Rephrase and expand the user's query as best as you can in one short sentence.
Query: < query >

$GenQCAR$: Using the definitions below, rephrase and expand the user's query as best as you can in one short sentence.
Synergy: Positive interactions where progress of an SDG promotes, contributes, or enables progress on another one.
Trade-off: Negative interactions where progress of an SDG deteriorates progress

towards another one.
Target < target_source >:
< target_source_definition >
Target < target_destination >:
< target_destination_definition >

C. Examples of query reformulation

q: *Are there synergies or trade-offs between SDG targets 17.11 and 10.7 ?*

$GenQ$: *What are the interrelations and potential benefits or costs of advancing targets 10.7 on sustainable and inclusive fiscal policies for development and 17.11 on effective public financial management for achieving the Sustainable Development Goals?*

$GenQCAR$: *How do progress on SDG target 17.11, which aims to significantly increase exports from developing countries, and target 10.7, which focuses on facilitating orderly migration, interact? Do they positively reinforce each other (synergy) or does progress on one target have a negative impact on the other (trade-off)?*

D. Augmented prompt to distinguish interlinkages

You are a text classifier. Your task is to analyse context intent and classify the context into SYNERGY or TRADE-OFF based on definitions below.
SYNERGY: Positive interactions where progress of an SDG promotes or enables progress on another one.
TRADE-OFF: Negative interactions where progress of an SDG deteriorates progress towards another one.
Analyze the given context and indicate whether it better illustrates SYNERGY or TRADE-OFF, justifying your response and explain your reasoning.
CONTEXT: < most_relevant_context >

E. Classification and underlying reasoning

$Mistral$: *Based on the context provided, it appears that SDG Water Sanitation and the target in question have a significant number of direct interlinkages, meaning that progress on this target contributes positively to the progress of SDG Water Sanitation. Additionally, the target in question is the direct driver for change for SDG Water Sanitation, as stated by "Total Direct Driver linkages deriving from Total Direct Response linkages going to 34." Furthermore, the target in question is indirectly linked with several other targets, including SDG 11, SDG 14 and SDG 15. While there may be some negative interactions or trade-offs between these targets and the target in question, the overall relationship between SDG Water Sanitation and the target in question is one of synergy. Therefore, based on the information provided, the context can be classified as SYNERGY.*