# Automatic GRI-SDG Annotation and LLM-Based Filtering for Sustainability Reports

Seyed Alireza Mousavian Anaraki[1,†], Danilo Croce[1,*,†] and Roberto Basili[1,†]

[1]Department of Enterprise Engineering, University of Rome Tor Vergata, Via del Politecnico 1, 00133, Rome, Italy

**Abstract**

Sustainability reports are often aligned with frameworks such as the Global Reporting Initiative (GRI) and the Sustainable Development Goals (SDGs), but large-scale, paragraph-level annotation remains a challenge. This paper introduces a fully automated pipeline that generates weak supervision by linking report paragraphs to GRI and SDG categories using structured content indices, official GRI-SDG mappings, and semantic similarity scoring. To mitigate the noise inherent in automatic annotation, we employ an instruction-tuned large language model (LLaMA 3.1) to filter assigned labels based on paragraph relevance. We evaluate the quality of our annotations through downstream SDG classification tasks on the OSDG Community Dataset, showing that LLM-based filtering aligns closely with human consensus and significantly improves model performance. Our results demonstrate that combining pruned, automatically annotated data with human-labeled examples leads to more accurate and robust SDG classification, supporting scalable, interpretable sustainability analysis.

**Keywords**

Sustainability Reporting, Sustainable Development Goals, Global Reporting Initiative, Large Language Models

## 1. Introduction

As the demand for transparent and accountable sustainability reporting continues to grow, organizations are increasingly expected to align their disclosures with well-established frameworks such as the Sustainable Development Goals (SDGs) [1], Global Reporting Initiative (GRI) [2], and Environmental, Social, and Governance (ESG) [3].

These frameworks provide the foundation for consistent and comparable sustainability metrics across sectors. However, sustainability reports are typically lengthy, unstructured PDF documents that blend qualitative narratives with quantitative data, making it challenging to extract meaningful insights, particularly at scale [4].

At the same time, the rise of Large Language Models (LLMs) has opened new avenues for automating and improving the quality of sustainability reporting. From extracting structured information to verifying claims and detecting inconsistencies, LLMs are now central to advancing natural language processing in this domain [5].

Annotating sustainability reports with SDG and GRI labels is essential for enabling downstream tasks such as benchmarking, automated scoring, and document classification. Structured annotation also facilitates cross-document analysis by aligning content across diverse reports and organizations.

Public efforts like the OSDG Community Dataset[1] provide valuable manual SDG annotations for policy documents and publication abstracts [6]; however, these resources remain limited in scope and are expensive to expand.

Recent work has addressed the limitations of manual sustainability annotation by developing automatic methods for labeling texts with SDG, GRI, and ESG categories [7, 2]. Building on this line of research, we propose an unsupervised annotation pipeline aimed at reducing both the cost and subjectivity of manual labeling. Our approach leverages GRI content indices, which serve as structured metadata in sustainability reports, linking disclosure topics to specific pages [8]. While these indices provide page-level associations for GRI standards, the actual correspondence at the paragraph level remains unknown; furthermore, we also seek to associate relevant SDG categories with each paragraph.

For example, consider the following excerpt from Merck's recent sustainability report: "*We promote equality, fairness, inclusion, and tolerance in the workplace by participating in initiatives such as the UN Women's Empowerment Principles and UN Global Compact's Target Gender Equality Programme.*" Through our pipeline, this paragraph can be automatically linked to the following categories:

- **SDG 5** (Gender): "*Achieve gender equality and empower women.*"
- **GRI 405** (Diversity and Equal Opportunity), specifically disclosure **GRI 405-2**: "*Ratio of basic*

[1]https://github.com/osdg-ai/osdg-data

*CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy*

*Corresponding author.

†These authors contributed equally.

seyedalireza.mousaviananaraki@students.uniroma2.eu (S. A. Mousavian Anaraki); croce@info.uniroma2.it (D. Croce); basili@info.uniroma2.it (R. Basili)

0009-0007-1044-9978 (S. A. Mousavian Anaraki); 0000-0001-9111-1950 (D. Croce); 0000-0001-5140-0694 (R. Basili)

*salary and remuneration of women to men."*

This example illustrates how individual report paragraphs can be meaningfully aligned with both the SDG and GRI frameworks; however, performing this mapping at scale is non-trivial. The full task involves 17 SDGs and 33 GRI standard codes (each with multiple disclosures), yielding hundreds of potential (GRI, SDG) combinations and significant ambiguity in narrative text. Addressing this challenge requires a systematic approach that can constrain the search space while preserving semantic relevance.

Our method bridges the gap between structured sustainability frameworks and unstructured report narratives, enabling large-scale and systematic annotation of disclosures. Concretely, we restrict the annotation search space by focusing on report pages linked to GRI standards in the content index, and further constrain possible annotations using established mappings between GRI codes and SDGs. This substantially reduces ambiguity and the combinatorial complexity inherent in considering all possible code pairs. To assign labels at the paragraph level, we compute semantic similarity between each paragraph and the textual definitions of GRI disclosures and SDG targets, using pre-trained sentence encoders [9, 10, 11]. This allows us to rank and select the most plausible (GRI, SDG) annotation pairs, resulting in a high-confidence, automatically annotated dataset.

Despite these constraints, unsupervised annotation methods—especially those based on bootstrapping and semantic similarity—can introduce noisy or weakly aligned labels. To address this, we propose a pruning strategy that further refines annotation quality. Specifically, we employ an instruction-tuned large language model (LLM), such as LLaMA 3.1 [12], to assess the contextual fit of each paragraph-label pair. The model is prompted to answer, in a binary fashion, whether the proposed annotation is relevant to the given paragraph. This step filters out misaligned pairs and improves the reliability of the final dataset for downstream sustainability analysis. While our implementation uses LLaMA 3.1, the approach is compatible with other instruction-tuned LLMs.

Directly assessing the quality of unsupervised annotations is inherently challenging due to the lack of ground-truth labels at scale. To address this, we adopt an indirect evaluation strategy: we train a supervised classifier on our pruned automatically annotated dataset and assess its performance on a well-established benchmark, the OSDG Community Dataset [6]. Our working hypothesis is that if the inclusion of pruned automatically annotated data leads to improved classification performance on the OSDG benchmark, then these data contribute useful information.[2] Preliminary results confirm that supplementing human-annotated data with pruned automatically annotated examples consistently improves classification accuracy, particularly for challenging or ambiguous texts.

We further evaluate the effectiveness of our pruning strategy through two complementary analyses. First, we leverage the structure of the OSDG Community Dataset, in which each text is associated not only with an SDG label but also with an agreement score, reflecting the proportion of annotators who endorsed the assigned label. By applying our LLM-based filtering method to OSDG, we examine the correlation between human consensus and the LLM's filtering decisions. Intuitively, a reliable pruning system should tend to retain annotations with high human agreement and filter more aggressively when annotator consensus is low, as these instances are more likely to be ambiguous or noisy. Our results show a clear alignment: paragraphs with high agreement scores are more frequently retained, while those with lower consensus are more likely to be discarded. Inspired by this analysis, we also examine the pruning behavior on automatically annotated data. We find a consistent trend: as the semantic similarity between a paragraph and its paired GRI-SDG labels increases, a larger proportion of annotations is retained. This suggests that LLaMA's filtering decisions are guided by semantic alignment, reinforcing the effectiveness of our similarity-based scoring approach for assessing label relevance.

Second, we directly compare downstream performance when training models on data with and without LLM-based filtering. Across all configurations, we observe that pruning improves overall classification accuracy. These findings suggest that the pruning step not only aligns with human judgments but also consistently enhances the utility of the resulting training data for sustainability text classification.

The remainder of this paper is organized as follows: Section 2 reviews the relevant literature. Section 3 introduces our automatic annotating and pruning methodology. Section 4 outlines the experimental setup and presents our evaluation results. Finally, Section 5 concludes the paper and discusses directions for future research.

## 2. Related Work

**Sustainability Reporting Frameworks.** Sustainability reporting is increasingly guided by global frameworks such as the United Nations Sustainable Development Goals (SDGs) [1], the Global Reporting Initiative (GRI)[3], and Environmental, Social, and Governance (ESG) principles [3]. The 2030 Agenda outlines 17 SDGs and 169 targets addressing major global development chal-

[2] Although our method generates both SDG and GRI labels, we focus on SDG evaluation in this work. Joint assessment of SDG and GRI annotations is left for future research.

[3] https://www.globalreporting.org/standards/

lenges [13], while the GRI, established in 1997, offers a structured framework for reporting economic, environmental, and social impacts [2]. It provides standardized disclosures—both required and recommended—that help organizations systematically communicate their contributions. To support SDG integration, the Action Platform Reporting on the SDGs[4], in collaboration with GRI, offers a database that maps SDG targets to specific GRI codes and disclosures, enabling companies to identify relevant reporting items and align strategic goals with operational metrics.

**Large Language Models in Sustainability Reporting** Large Language Models (LLMs) have become powerful tools in natural language processing, offering innovative solutions to longstanding challenges in sustainability reporting. Their high accuracy and adaptability make them well-suited for extracting structured data, performing textual analysis, and identifying misleading green claims [5].

LLMs are typically categorized into three main types based on their neural architecture: encoder-only, decoder-only, and encoder-decoder models [14].

**Encoder-only models**, such as BERT [15], focus on encoding the input text into rich contextual representations using self-attention mechanisms. These models are especially effective for classification and interpretive tasks like sentiment analysis and named entity recognition. These models dominate sustainability NLP applications due to their high performance on classification tasks. They have been widely used for aligning corporate texts with SDGs [16, 17, 18], GRI [19], and ESG [20, 21, 22]. Models like BERT, RoBERTa, SBERT, MiniLM, and DistilBERT are frequently fine-tuned to extract structured insights and detect misleading green claims using ClimateBERT [23] and MacBERT [24]. For example, ESG-KIBERT [20] employs an encoder-only architecture specifically designed for industry-specific ESG evaluation, demonstrating how domain adaptation can improve the performance of deep language models in sustainability contexts.

**Decoder-only models**, such as LLaMA [12], operate auto-regressively by predicting one token at a time conditioned on prior outputs. This makes them suitable for generative tasks such as text completion, summarization, and dialogue generation. Recent studies underscore the growing role of decoder-only models in sustainability reporting, particularly through their integration with retrieval-augmented generation (RAG) techniques [25], as demonstrated in ESG applications by Bronzini et al. [26] and Zou et al. [3]. Additionally, Jain et al. [27] highlighted the effectiveness of GPT-3.5 in addressing ESG-related prompts and identifying nuanced sustainability issues.

**Encoder-decoder models** like BART [28] combine text understanding and generation, making them well-suited for complex tasks such as summarization. Though less commonly used, they have proven effective in sustainability reporting—e.g., BART was used for SDG multi-label categorization [29].

Following the trends outlined above, our approach assigns task-specific roles to decoder-only and encoder-only LLMs based on their architectural strengths. We use LLaMA 3.1—an instruction-tuned decoder-only model—to filter noisy or weakly aligned GRI-SDG annotations through generative prompting, guided by an embedding-based similarity scoring process. Specifically, we use a pre-trained MpNet model to compute alignment scores between each paragraph and its associated GRI-SDG label descriptions, allowing us to generate more semantically grounded annotations by prioritizing label pairs with the highest similarity. For downstream classification, we fine-tune a BERT-based encoder model for multi-label SDG prediction, capitalizing on its effectiveness in structured, discriminative tasks. This design reflects a practical alignment between model capabilities and task requirements in the context of sustainability reporting. Moreover, by improving the quality of both human and automatically annotated data, our approach contributes to more reliable alignment with established reporting standards such as the SDGs and GRI, thereby supporting more transparent and accountable sustainability disclosures.

## 3. Automatic Paragraph Annotation via Structured Indices, Semantic Similarity, and LLM Filtering

We present a multi-step pipeline for automatically annotating paragraphs from sustainability reports with both GRI (Global Reporting Initiative) and SDG (Sustainable Development Goals) labels. The process leverages document structure, official mappings, and semantic similarity, with a final human-like filter based on a large language model.

**Paragraph Segmentation and Preprocessing.** Each report is parsed with a layout-aware tool (e.g., PyMuPDF[5]), extracting all text blocks and filtering out headers, footers, and fragments. Only blocks of at least 20 words are retained as candidate paragraphs.

[4] https://www.globalreporting.org/reporting-support/goals-and-targets-database/

[5] https://github.com/pymupdf/PyMuPDF

For example, a typical extracted paragraph might be: *"In 2023, CompanyX reduced its greenhouse gas emissions by 15% by switching to renewable energy sources. The organization remains committed to transparent reporting of its climate targets and actions."*

**Generating Candidate and Alternative Labels.** Most reports include a *GRI content index*, a table authored by the company that indicates, for each GRI disclosure code (e.g., GRI 305: Emissions, GRI 302: Energy), the specific pages where the disclosure is addressed.

For each paragraph $p$ occurring on page $\pi$, we define:

- The **candidate set** as all GRI codes explicitly linked to $\pi$ via the content index.
- The **alternative set** as all remaining GRI codes not mentioned in the index for $\pi$, but potentially relevant based on semantic content.

Continuing the example, suppose the GRI content index indicates that the pages containing the paragraph above refer to GRI 305 (Emissions) and GRI 302 (Energy). These two codes are included in the *candidate set* for the paragraph, as they are explicitly claimed by the report on that page. All remaining GRI codes—among the approximately 33 topical standards defined in the GRI framework—are considered part of the *alternative set*. These alternatives are not mentioned in the content index for this page, but may still be semantically relevant to the paragraph based on its content. Note that, due to the broad and multi-faceted nature of sustainability topics, the content index is not expected to capture all relevant GRI standards for each page. It typically highlights the main disclosures, while secondary or nuanced themes may be omitted. By considering both the candidate set (directly indexed codes) and the alternative set (other potentially relevant codes), our approach accounts for both explicit priorities and additional associations present in the narrative.

**Expansion to SDG Pairs via Official Mapping.** Each GRI code captures a specific disclosure standard (e.g., energy consumption, gender pay equality), while each SDG describes a broader societal goal (e.g., SDG 7: Affordable and Clean Energy; SDG 5: Gender Equality). To bridge these conceptual levels in a principled way, we use the official mapping[6] $\mathcal{M}$, which links each GRI code only to semantically relevant SDG targets.

This mapping is essential for two reasons: (i) it avoids generating irrelevant or misleading (GRI, SDG) pairs—since not every combination is meaningful in practice (e.g., GRI 305: Emissions is unrelated to SDG 4: Quality Education)—and (ii) it guarantees that downstream semantic similarity scoring is only performed between a paragraph and label pairs with a recognized conceptual connection, thus improving interpretability and actionability for sustainability analysis.

Given a paragraph $p$, we use its associated GRI codes—those directly referenced in the content index (candidate set) and all other codes not mentioned (alternative set)—to generate all valid triples $(p, g, s)$, where $s \in \mathcal{M}(g)$. For example, as above:

- GRI 305 maps to SDG 13 (Climate Action),
- GRI 302 maps to both SDG 13 and SDG 7 (Affordable and Clean Energy).

This produces two filtered sets of candidate triples: those based on content-indexed GRI codes, and those based on alternative codes. For the running example, the triples derived from the content index are:

- (paragraph, GRI 305, SDG 13),
- (paragraph, GRI 302, SDG 13),
- (paragraph, GRI 302, SDG 7).

At this stage, all generated triples are semantically plausible and ready for embedding-based similarity scoring.

**Semantic Similarity Ranking.** Even after filtering out irrelevant combinations via the official GRI→SDG mapping, each paragraph remains associated with a large number of possible label pairs. We therefore rank all remaining (paragraph, GRI, SDG) triples based on how semantically aligned they are with the paragraph content.

To quantify alignment, we use a pre-trained sentence encoder (MPNet [9]) to compute cosine similarities in embedding space. For each triple, we consider the textual description of the SDG target and all available disclosure requirements associated with the GRI code. We define the similarity score $\sigma(p, g, s)$ as:

$$\sigma(p, g, s) = \max_{r \in R_g} \cos(\mathbf{e}_p, \mathbf{e}_r) \cdot \max_{t \in T_s} \cos(\mathbf{e}_p, \mathbf{e}_t)$$

where $\mathbf{e}_p$ is the embedding of the paragraph, $R_g$ is the set of disclosure texts for GRI code $g$, and $T_s$ is the set of textual definitions for SDG $s$ (typically the goal and its targets). This formulation favors pairs for which both components—GRI and SDG—are independently relevant to the paragraph: if either component is weakly aligned, the product score will be low. This reflects the intuition that a good annotation should simultaneously satisfy both frameworks. For example, suppose a paragraph discusses emissions reduction due to renewable energy adoption. We obtain:

- $\cos(\text{paragraph}, \text{GRI 305}) = 0.92$ (strong match with "Reduction of GHG emissions"),

[6] https://www.globalreporting.org/reporting-support/goals-and-targets-database/

- $\cos(\texttt{paragraph}, \texttt{SDG 13}) = 0.88$ (climate action),
- $\cos(\texttt{paragraph}, \texttt{GRI 302}) = 0.69$ (energy reduction consumption),
- $\cos(\texttt{paragraph}, \texttt{SDG 7}) = 0.54$ (clean energy).

The resulting joint scores are: (GRI 305, SDG 13): $0.92 \times 0.88 = 0.81$, (GRI 302, SDG 13): $0.69 \times 0.88 = 0.61$, (GRI 302, SDG 7): $0.69 \times 0.54 = 0.37$.

Notably, we compute these scores for both candidate and alternative triples. While candidate triples originate from the GRI content index (i.e., the report explicitly claims these topics are discussed on the page), alternative triples arise from GRI codes not mentioned in the index. Though potentially less reliable, alternative labels may capture omissions or relevant but unindexed content. Hence, if a triple from the alternative set obtains a substantially higher semantic score than those in the candidate set, it may signal that the original index missed something. In this case, our strategy allows the model to retain the best alternative triple. While semantic similarity offers a useful initial filter, it may miss deeper context or introduce noise. To address this, we add later an LLM-based filtering step for more robust alignment.

**Disambiguation Policies: Conservative and Permissive.** After ranking all (paragraph, GRI, SDG) triples by joint semantic similarity, the final step is to select which annotations to retain for each paragraph. This choice must balance precision (avoiding spurious labels) with recall (capturing genuine but possibly under-indexed content). We propose two complementary disambiguation policies, which reflect different trade-offs between coverage and selectivity.

**Conservative Policy:** This policy is tailored for high-precision applications, where false positives are especially costly. For each paragraph, we:

1. Identify the best-scoring candidate triple (i.e., derived from the GRI codes listed in the report's index for the relevant page).
2. Identify the best-scoring alternative triple (i.e., derived from any other valid (GRI, SDG) pair for the paragraph).
3. If the candidate triple's score is greater than or equal to the alternative's, we retain only the candidate triple—reflecting high confidence in the company's index.
4. If the alternative triple has a higher score, we return both the best candidate and the best alternative. This accounts for possible omissions or underreporting in the index, while maintaining interpretability.

In practice, this policy outputs either one or two annotation triples per paragraph.

**Permissive Policy:** This policy is designed to maximize recall and accommodate semantic ambiguity—useful for exploratory analysis or downstream expert curation.

1. Find the candidate triple with the highest score and set a threshold at half that value.
2. Retain up to two candidate triples whose scores exceed this threshold (to account for ties or near-equivalent topics).
3. Always include the best-scoring alternative triple, regardless of its absolute score, ensuring that strong semantic signals outside the index are never discarded a priori.

As a result, this policy can return up to three triples (two candidates plus one alternative) for a given paragraph, allowing for richer, multi-label annotation. In summary, the conservative policy favors precision, whereas the permissive policy promotes recall and label diversity.

**Final Filtering with LLM Relevance Assessment** While semantic similarity models are powerful for linking text to structured concepts, they can sometimes overestimate relevance—especially for vague, generic, or multi-topic paragraphs. For example, a paragraph mentioning "sustainable growth" could weakly match almost any SDG, leading to noisy or spurious labels even after careful mapping and scoring.

To further improve annotation quality, we add a final "human-like" relevance check using a large language model (LLM) such as LLaMA 3.1 Instruct. This step serves two key purposes: i) it filters out weak, contextually inappropriate, or overly broad matches that the similarity-based method might miss; ii) it simulates expert review at scale, bringing richer contextual understanding and nuanced judgment—skills typically seen in human annotators—while maintaining automation and consistency.

For each retained (paragraph, GRI, SDG) triple, we construct a structured prompt (shown in Figure 1) presenting the paragraph and the official descriptions of both labels. The LLM is asked to answer—based solely on the evidence given—whether the label pair is truly relevant to the paragraph content. Only those triples receiving a "Yes" are included in the final dataset.

For instance, a paragraph describing the company's general commitment to "sustainable development" might weakly match several SDGs and GRIs in embedding space, but only a focused LLM assessment can determine if a specific (GRI, SDG) pair is truly justified by the text. In this way, the LLM acts as a high-precision, scalable expert-in-the-loop filter. This LLM-based filtering step significantly reduces false positives, capturing complex connections and subtle mismatches that even strong embedding models may overlook. In effect, it combines the scale and speed of automated annotation with the contextual depth

You are a sustainability evaluation assistant. Decide if the following GRI–SDG pair is relevant to the paragraph.

**Paragraph:** "Paragraph content here"

**GRI [GRI Code]:** GRI Description here
**SDG [SDG Name]:** SDG Description here

Only reply with one word: **Yes** or **No**.

**Format:**
Answer: Yes
(or)
Answer: No

**Figure 1:** LLM prompt for paragraph-level GRI–SDG relevance filtering. The model is asked to decide, given the paragraph and both label descriptions, if the label pair is truly relevant. Only a one-word response (**Yes** or **No**) is permitted.

of human reasoning, resulting in a cleaner, more trustworthy annotated dataset ready for downstream analysis or model training.

## 4. Experimental Evaluation

We conduct a comprehensive experimental evaluation to assess the effectiveness of our automatic annotation pipeline and its LLM-based filtering component. Our analysis focuses on two main questions: (i) does LLM filtering produce label decisions that align with human consensus? and (ii) how do different label selection policies (conservative vs. permissive) and LLM filtering impact the quality and utility of the resulting annotated data for downstream SDG classification?

### 4.1. LLM Filtering and Human Consensus on OSDG-CD

A natural concern when introducing LLM-based filtering into any annotation pipeline is whether the model's binary "Yes/No" relevance judgments are in fact consistent with human annotation practices. While LLMs are increasingly adopted as automated evaluators or assistants, there is limited empirical evidence on how closely their filtering behavior tracks with actual human agreement—particularly in specialized domains such as sustainability. To address this, we leverage the OSDG Community Dataset (OSDG-CD), a large-scale benchmark in which each paragraph-SDG pair is annotated not only with the assigned label, but also with an explicit agreement score reflecting the proportion of human annotators who supported the label assignment. This agreement score provides a direct, interpretable measure of human consensus, ranging from 0.1 (highly ambiguous or disputed cases) to 1.0 (full agreement among annotators). We use the LLaMA 3.1 Instruct model as a post-hoc filter: for each paragraph-SDG pair in OSDG-CD, we prompt the model to decide if the label is relevant to the paragraph, using the same structured format adopted in our main pipeline. We then analyze the fraction of examples retained ("Yes" by the LLM) across different agreement intervals.

**Table 1**
Distribution of agreement scores in OSDG-CD.

| Agreement Interval | Frequency |
|---|---|
| [0.1, 0.3) | 2,321 |
| [0.3, 0.5) | 5,249 |
| [0.5, 0.7) | 7,064 |
| [0.7, 1) | 6,041 |
| 1.0 | 14,922 |

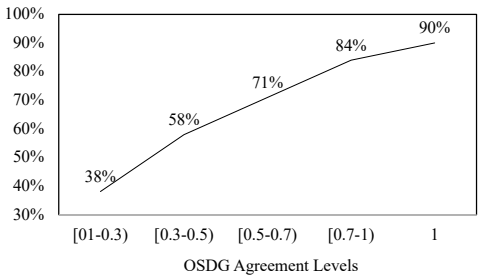


**Figure 2:** LLM filtering aligns with human agreement: retention rates ("Yes" answers) by the LLaMA 3.1 model increase with human consensus in the OSDG-CD dataset.

Table 1 reports the frequency distribution of samples across agreement bins, and Figure 2 visualizes the key result: the likelihood of a sample being retained by the LLM filter increases monotonically with human agreement. In other words, pairs with high human consensus are almost always preserved by the model, while those with low or disputed agreement are more frequently filtered out. This positive correlation provides strong evidence that LLM-based filtering is not arbitrary, but instead captures a notion of relevance that closely mirrors collective human judgment.

This result has two important implications. First, it provides empirical support for using LLMs as scalable, "expert-in-the-loop" filters for semantic annotation, even in cases where manual adjudication would be prohibitively expensive. Second, it suggests that LLMs can help mitigate annotation noise in weakly or ambiguously labeled data—removing many of the examples that humans themselves would likely judge as borderline or unreliable. Overall, this agreement-guided analysis not only validates our specific use of LLM filtering in the

construction of GRI-SDG training data, but also suggests a broader role for LLMs as automatic quality controllers in human-in-the-loop NLP pipelines.

## 4.2. Assessing Labeling Strategies for Automatic Paragraph Annotation

**Experimental Setup.** To systematically evaluate our annotation pipeline, we applied it to a curated corpus of 30 sustainability reports spanning 10 sectors and 3,663 pages. After preprocessing and paragraph segmentation, we obtained 19,133 candidate paragraphs, of which 10,303 were indexed by company-provided GRI content indices and thus eligible for annotation. Annotation followed the multi-step procedure described in Section 3: we generated (GRI, SDG) label pairs using the official mapping, scored their semantic similarity, and selected final annotations according to either the conservative (high-precision, at most one or two triples per paragraph) or permissive (higher recall, up to three triples) policy.

Applying the conservative policy yielded 17,216 label pairs initially, which were reduced to 4,558 after LLM-based relevance filtering. The permissive policy produced a higher initial volume of annotations (30,647 label pairs), which was pruned to 7,425 after filtering with LLaMA 3.1 Instruct. This substantial reduction confirms the impact of the LLM-based step in filtering out weak or noisy annotations, ultimately improving the quality and reliability of the final labeled dataset. For evaluation, we leveraged the OSDG Community Dataset (OSDG-CD), which contains single-label SDG assignments per paragraph, validated by crowdsourced agreement scores. To ensure reliability, we defined two test splits: a **Simple** set (agreement = 1.0, fully unambiguous) and a **Complex** set ($0.7 \leq$ agreement $\leq 1.0$). All models were trained in a multi-label setting, but evaluated using only the highest-scoring prediction per paragraph to match the OSDG single-label ground truth. As a baseline, we used a BERT-based classifier (`bert-base-cased`). We used a standard binary cross-entropy loss for multi-label classification over the full label set, treating each label independently during training. The model was trained with an effective batch size of 16 (via gradient accumulation over 4 mini-batches of size 4), using the AdamW optimizer with a learning rate of $2 \times 10^{-5}$, weight decay of 0.1, and a linear learning rate scheduler with a warmup ratio of 0.1, for a total of 5 training epochs. Accuracy is defined as the percentage of paragraphs for which the top predicted label matches the ground truth; since the OSDG test set provides only one true label per paragraph, this top-1 accuracy measure is equivalent to precision, recall, and F1-score, which are therefore omitted.

**Does LLM Filtering Improve Automatically Annotated Training Data?** Our first experiment tests whether LLM-based filtering effectively improves the utility of automatically annotated data, and how the choice of annotation policy (conservative vs. permissive) impacts downstream model performance.

**Table 2**
Accuracy on OSDG test sets with different training sets: conservative vs permissive policy, before and after LLM filtering.

| Training | Simple | Complex |
|---|---|---|
| Conservative | 0.762 | 0.737 |
| Conservative + LLM | **0.783** | **0.752** |
| Permissive | 0.688 | 0.660 |
| Permissive + LLM | **0.726** | **0.695** |

Results (Table 2) indicate that both policies benefit from LLM filtering, but to different extents. The conservative policy (high-precision, fewer labels) already yields reasonably strong results, but applying LLM filtering further increases accuracy by removing residual false positives. The permissive policy (higher recall, more candidate triples per paragraph) initially introduces substantially more noise, as reflected in lower baseline accuracy; however, LLM filtering provides a larger relative improvement—yet, even after filtering, the permissive setting still lags behind the conservative one in absolute performance. This suggests that, while the LLM can mitigate a large portion of annotation noise, excessive over-labeling (as in the permissive setting) cannot be fully corrected in post-processing, and some spurious associations may persist. In summary, LLM-based filtering systematically improves the quality of automatically generated labels, especially in the presence of noisy or overly broad candidate assignments. However, the conservative policy remains preferable in settings where downstream precision is paramount[7].

**Does Adding Automatically Annotated Data Benefit Supervised Training?** In a second experiment, we assessed whether supplementing human-annotated data (OSDG-CD) with LLM-pruned automatic annotations yields tangible improvements in SDG classification.

**Table 3**
Accuracy on OSDG test sets with and without adding pruned automatic data (Cons.: conservative, Perm.: permissive).

| Training | Simple | Complex |
|---|---|---|
| OSDG (full) | 0.917 | 0.907 |
| OSDG + Cons. + LLM | **0.921** | **0.910** |
| OSDG + Perm. + LLM | 0.919 | 0.909 |

[7] Note that the test set requires a single SDG per paragraph, so we evaluate our classifier by selecting only the top prediction. This may not capture all relevant SDGs, especially for complex cases, but gives a reasonable first estimate of performance.

Results in Table 3 show that, for both policies, adding pruned automatic annotations to the OSDG training set consistently increases accuracy on both simple and complex test splits. While the gains are modest, they are robust across settings, confirming that our pipeline produces useful complementary signal even in the presence of expert-labeled data. As in the previous experiment, the conservative policy remains more reliable, providing slightly higher accuracy than the permissive policy; the latter, despite contributing more examples, appears to introduce a small amount of residual noise that is not fully eliminated by LLM filtering.

Taken together, these findings support a dual conclusion: (1) the automatic annotation pipeline is effective for scalable SDG data generation, and (2) the interplay between label selection policy and LLM-based filtering is crucial for balancing coverage and precision. The conservative strategy, enhanced by LLM filtering, delivers high-quality labels that boost supervised learning, while the permissive strategy is valuable for recall-oriented applications but requires careful calibration to avoid excessive noise.

## 4.3. Analysis of LLM Retention Decisions on Automatically Annotated Data

Having established that the LLM-based filter is well aligned with human consensus on the OSDG dataset (Section 4.1), we next analyze how the LLM's binary relevance judgments interact with the underlying semantic similarity scores in our full, automatically annotated dataset. This provides a deeper understanding of whether the LLM filter simply introduces an arbitrary bottleneck, or if it systematically reinforces semantic quality.

We consider the *product similarity score*—the product of cosine similarities between a paragraph and its associated GRI and SDG descriptions (see Section 3)—as a measure of semantic alignment for each candidate label. For every (paragraph, GRI, SDG) triple, we record whether the LLM filter retained the annotation ("Yes") or discarded it ("No"). Table 4 reports the mean similarity scores for retained and discarded samples, disaggregated by both label type (Candidate, Alternative) and selection policy (Conservative, Permissive).

As shown, the LLM filter systematically prefers to retain labels with higher semantic similarity to the paragraph, regardless of whether they are candidate or alternative labels, and across both policies. The effect is particularly pronounced for alternatives, which are only kept when they exhibit a strong semantic match.

To further examine this relationship, we discretize the similarity scores into bins and calculate, for each bin, the proportion of samples retained by the LLM. Figure 3 presents these retention rates for the conservative (Top-1) policy, separately for candidates, alternatives, and the combined set. To ensure statistical significance, we only report bins containing at least 700 samples. The threshold of 700 samples was chosen empirically based on the distribution of paragraph counts across prediction score intervals. Specifically, we observed that the total number of samples in the higher-confidence intervals—i.e., those greater than 0.7 ((0.7-0.8], (0.8-0.9], (0.9-1])—was only 272 (227 + 40 + 5). Given such low sample sizes, reporting performance metrics for these bins would risk statistical instability and lack of representativeness. To mitigate this, we selected 700 as a minimum cutoff to ensure that each bin included in our analysis contains a sufficient number of samples for reliable metric estimation. This threshold balances coverage across confidence intervals with the statistical reliability of the reported results.

**Table 4**
Mean product similarity score for retained vs. discarded samples under conservative and permissive label selection.

| Policy | Category | Retained | Discarded |
|---|---|---|---|
| Conservative | Overall | 0.434 | 0.321 |
| | Alternatives | 0.463 | 0.351 |
| | Candidates | 0.422 | 0.298 |
| Permissive | Overall | 0.414 | 0.308 |
| | Alternatives | 0.456 | 0.353 |
| | Candidates | 0.400 | 0.283 |

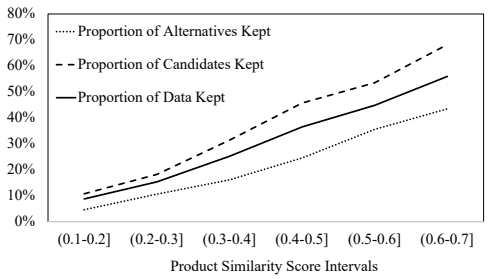


**Figure 3:** Proportion of (paragraph, GRI, SDG) triples retained by the LLM filter as a function of product similarity score, binned by intervals. Results are shown for candidate, alternative, and all labels under the conservative (Top-1) policy.

The figure demonstrates a clear monotonic trend: as the product similarity score increases, the probability of retention by the LLM rises sharply. For scores below 0.3, fewer than 20% of labels are retained, while for scores above 0.6, the retention rate exceeds 60%. This pattern holds for both candidates and alternatives, further supporting the conclusion that the LLM acts as a semantic relevance filter—amplifying the selectivity of the automatic annotation pipeline and systematically favoring labels with strong textual alignment.

In summary, these results indicate that our LLM-based filtering mechanism is not merely an arbitrary post-processing step, but an effective semantic validator: it consistently prioritizes label assignments with robust evidence in the paragraph text.

## 5. Conclusion and Future Work

This work presents a fully automated pipeline for large-scale annotation of sustainability reports at paragraph level, aligning text with both GRI disclosures and SDG targets. Leveraging structured metadata, official GRI-SDG mappings, semantic similarity, and an LLM-based relevance filter (LLaMA), our method offers an interpretable and scalable alternative to manual annotation. The LLM filter proves highly effective in reducing semantic noise and producing annotations that closely match human consensus.

Our experiments show that LLaMA-based filtering favors labels with high semantic similarity, aligns with human judgments on the OSDG benchmark, and consistently improves downstream SDG classification—even when combined with expert-labeled data. While permissive labeling increases coverage, it also adds noise that is only partly corrected by LLM filtering.

This pipeline lays the foundation for more transparent and data-driven sustainability analytics. Future research will focus on several open challenges. First, we aim to expand the LLM filter to provide natural language justifications for its decisions, improving explainability and facilitating expert validation. We also acknowledge that scalability may become a limitation when applying our pipeline to thousands of reports, particularly due to the computational cost of LLM-based filtering; addressing this bottleneck through optimization or distillation techniques is a key direction for future work. Second, while our current evaluation is primarily model-based, we plan to conduct in-depth human studies, including manual validation of high-confidence (GRI, SDG) pairs, and direct comparisons with prior supervised approaches [16, 18], especially regarding the annotation of GRI codes. Third, we envision extending our framework to cover a wider array of sustainability and ESG standards, as well as to support fine-grained analysis of the substance and quality of sustainability reporting—such as distinguishing between specific, verifiable disclosures and generic statements, thus advancing automated detection of greenwashing.

## Acknowledgments

We thank Armando Calabrese, Roberta Costa and Luigi Tiburzi for their valuable advice, insightful discussions on sustainability reporting, and for generously sharing documents that inspired and enabled this initial experimental study. We acknowledge financial support from the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

## References

[1] V. Nationen, Transforming Our World: The 2030 Agenda for Sustainable Development: A/Res/70/1, United Nations, Division for Sustainable Development, 2015.

[2] H. Q. Ngee, A. Ganesh, M. A. N. Azmi, T. Y. Tang, M. Mukred, F. Mohammed, A. A. B. Ahmad, Environmental, social and governance (esg) scores automation in global reporting initiative (gri) with natural language processing, in: Proc. 2024 7th Int. Conf. Internet Appl., Protocols, and Services (NETAPPS), 2024, pp. 1–7.

[3] Y. Zou, M. Shi, Z. Chen, Z. Deng, Z. Lei, Z. Zeng, S. Yang, H. Tong, L. Xiao, W. Zhou, Esgreveal: An llm-based approach for extracting structured data from esg reports, J. Clean. Prod. 489 (2025) 144572.

[4] H. Kang, J. Kim, Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods, Appl. Sci. 12 (2022) 5614.

[5] W. Moodaley, A. Telukdarie, A conceptual framework for subdomain specific pre-training of large language models for green claim detection, Eur. J. Sustain. Dev. 12 (2023) 319. doi:10.14207/ejsd.2023.v12n4p319.

[6] L. Pukelis, N. Bautista-Puig, G. Statulevičiūtė, V. Stančiauskas, G. Dikmener, D. Akylbekova, Osdg 2.0: A multilingual tool for classifying text data by un sustainable development goals (sdgs), arXiv preprint abs/2211.11252 (2022). Available at: https://arxiv.org/abs/2211.11252.

[7] C. Jakob, V. Schmitt, S. Mohtaj, S. Möller, Classifying sustainability reports using companies self-assessments, in: Future of Information and Communication Conference, Springer, 2024, pp. 547–557.

[8] I. Nechaev, D. S. Hain, Social impacts reflected in csr reports: Method of extraction and link to firms' innovation capacity, J. Clean. Prod. 429 (2023) 139256.

[9] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, Adv. Neural Inf. Process. Syst. 33 (2020) 16857–16867.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang.

Technol. (NAACL-HLT), Vol. 1 (Long and Short Papers), 2019, pp. 4171–4186.

[11] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proc. 2019 Conf. Empirical Methods Nat. Lang. Process. and 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP), 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.

[12] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023). Available at: https://arxiv.org/abs/2302.13971.

[13] T. B. Smith, R. Vacca, L. Mantegazza, I. Capua, Natural language processing and network analysis provide novel insights on policy and scientific discourse around sustainable development goals, Sci. Rep. 11 (2021) 22427. doi:10.1038/s41598-021-01801-6.

[14] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: A survey, Sci. China Technol. Sci. 63 (2020) 1872–1897. doi:10.1007/s11431-020-1647-3.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018). Available at: https://arxiv.org/abs/1810.04805.

[16] M. Angin, B. Taşdemir, C. A. Yılmaz, G. Demiralp, M. Atay, P. Angin, G. Dikmener, A roberta approach for automated processing of sustainability reports, Sustain. 14 (2022) 16139. doi:10.3390/su142316139.

[17] Y. Li, M. Rockinger, Unfolding the transitions in sustainability reporting, Sustain. 16 (2024) 809. doi:10.3390/su16020809.

[18] R. Yao, M. Tian, C.-U. Lei, D. K. W. Chiu, Assigning multiple labels of sustainable development goals to open educational resources for sustainability education, Educ. Inf. Technol. 29 (2024) 18477–18499.

[19] L. Hillebrand, M. Pielka, D. Leonhard, T. Deußer, T. Dilmaghani, B. Kliem, R. Loitz, M. Morad, C. Temath, T. Bell, et al., sustain.ai: a recommender system to analyze sustainability reports, in: Proc. 19th Int. Conf. Artif. Intell. Law, 2023, pp. 412–416. doi:10.1145/3594536.3595131.

[20] H. Lee, J. H. Kim, H. S. Jung, Esg-kibert: A new paradigm in esg evaluation using nlp and industry-specific customization, Decis. Support Syst. 193 (2025) 114440.

[21] T. Schimanski, A. Reding, N. Reding, J. Bingler, M. Kraus, M. Leippold, Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication, Finance Res. Lett. 61 (2024) 104979. doi:10.1016/j.frl.2024.104979.

[22] A. Gupta, A. Chadha, V. Tewari, A natural language processing model on bert and yake technique for keyword extraction on sustainability reports, IEEE Access (2024). doi:10.1109/ACCESS.2024.3352742.

[23] A. Vinella, M. Capetz, R. Pattichis, C. Chance, R. Ghosh, Leveraging language models to detect greenwashing, arXiv preprint arXiv:2311.01469 (2023). Available at: https://arxiv.org/abs/2311.01469.

[24] X. Wang, X. Gao, M. Sun, Construction and analysis of corporate greenwashing index: a deep learning approach, EPJ Data Sci. 14 (2025) 1–25.

[25] K. Mehul, V. R. Kanagavalli, K. R. Saradha, P. N. Gowtham, M. P. Sachin, U. Surya, R. Godhandaraman, S. Girish, R. Naveen, Gen ai driven faq chatbot using advanced rag architecture for querying annual reports, in: Proc. 2025 Int. Conf. Comput. Commun. Technol. (ICCCT), 2025, pp. 1–6.

[26] M. Bronzini, C. Nicolini, B. Lepri, A. Passerini, J. Staiano, Glitter or gold? deriving structured insights from sustainability reports via large language models, EPJ Data Sci. 13 (2024) 41. doi:10.48550/arXiv.2310.05628.

[27] Y. Jain, S. Gupta, S. Yalciner, Y. N. Joglekar, P. Khetan, T. Zhang, Overcoming complexity in esg investing: The role of generative ai integration in identifying contextual esg factors, SSRN (2023). Available at SSRN 4495647.

[28] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019). Available at: https://arxiv.org/abs/1910.13461.

[29] F. Gonzalez, Z. Jin, B. Schölkopf, T. Hope, M. Sachan, R. Mihalcea, Beyond good intentions: Reporting the research landscape of nlp for social good, arXiv preprint arXiv:2305.05471 (2023). Available at: https://arxiv.org/abs/2305.05471.

## Online Resources

- OSDG Community Dataset,
- United Nations Sustainable Development Goals (SDGs)
- Global Reporting Initiative (GRI)
- GRI-SDG Mapping