Combatting Climate Change: Enhancing ESGLLM Retrieval with ESG-CID, a Disclosure Content Index Dataset for Mapping GRI and ESRS

Anonymous ACL submission

Abstract

Climate change has intensified the need for transparency and accountability in organizational practices, making Environmental, Social, and Governance (ESG) reporting increasingly crucial. Frameworks like the Global Reporting Initiative (GRI) and the new European Sustainability Reporting Standards (ESRS) aim to standardize ESG reporting, yet generating comprehensive reports remains challenging due to the considerable length of ESG documents and variability in company reporting styles. To facilitate ESG report automation, Retrieval-Augmented Generation (RAG) systems can be employed, but their development is hindered by a lack of labeled data suitable for training retrieval models. In this paper, we leverage an underutilized source of weak supervision—the disclosure content index found in past ESG reports—to create a comprehensive dataset, `ESG-CID`, for both GRI and ESRS standards. By extracting mappings between specific disclosure requirements and corresponding report sections, and refining them using a Large Language Model as a judge, we generate a robust training and evaluation set. We benchmark popular embedding models on this dataset and show that fine-tuning smaller BERT-based models can outperform commercial embeddings and leading public models, even under temporal data splits and cross-report style transfer from GRI to ESRS.

1 Introduction

Addressing climate change is one of the most pressing challenges of our time. This accelerating global climate crisis and increasing societal demands for corporate accountability have made Environmental, Social, and Governance (ESG) reporting a critical aspect of modern business. Natural Language Processing plays a pivotal role in understanding and drafting these documents. Recent advancements in Large Language Models (LLMs) enable the analysis of vast amounts of textual data related to climate

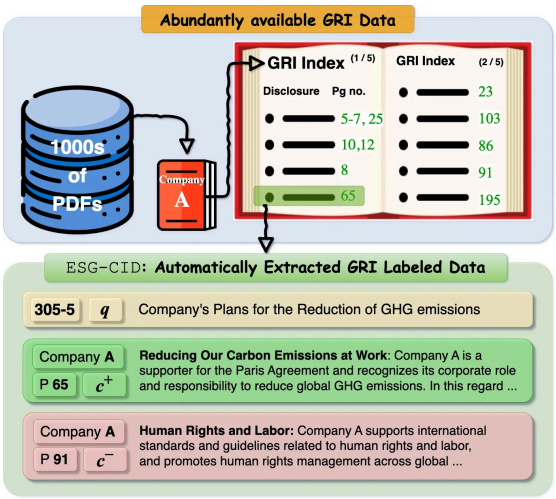

Figure 1: We extract content indices from GRI-compliant sustainability PDFs to create an ESG relevance dataset: ESG-CID. Each entry consists of a disclosure query (q), a relevant chunk (c^+) from the indexed page, and a randomly selected irrelevant chunk (c^-) from the rest of the document

policies, sustainability reports, and environmental impact assessments (Vaghefi et al., 2023; Schimanski et al., 2024). By extracting actionable insights from ESG reports, LLMs enhance transparency and inform stakeholders, driving data-driven decision-making in sustainability practices.

Despite these advancements, generating comprehensive and standardized ESG reports remains a significant challenge. ESG documents are extensive—averaging 120 pages—and exhibit variability in reporting styles and structures among organizations. The lack of standardized and accessible ESG data can lead to greenwashing, obscures true risks, and impedes the effective allocation of resources toward sustainable investments and practices. Frameworks like the Global Reporting Initiative (GRI) and the new European Sustainability Reporting Standards (ESRS) aim to standardize ESG reporting, but automating this process requires

effective Retrieval-Augmented Generation (RAG) systems. The development of such systems is hindered by a lack of labeled data suitable for training and evaluating retrieval models in the ESG domain.

The scarcity of labeled data arises mainly due to two factors: First, the considerable length of ESG reports makes manual annotation labor-intensive and time-consuming. Second, the lack of uniformity in reporting styles across different companies presents a challenge in creating datasets that generalize well. The combination of these factors makes it difficult to develop robust retrieval models needed for automating ESG reporting tasks.

In this paper, we leverage an underutilized yet readily available source of weak supervision: the **disclosure content index** found in past reports. We observed that GRI-compliant reports often include a content index linking specific disclosure requirements to corresponding sections or page numbers within the report. By extracting these mappings, we can generate large amounts of weakly supervised data that associates ESG disclosure queries with relevant text passages. To enhance the quality of this data, we employ a Large Language Model (LLM) as a judge to refine and validate the mappings. This process enables us to create a comprehensive dataset for both GRI and ESRS standards.

Using this dataset, we benchmark popular embedding models on the ESG retrieval task and explore the impact of fine-tuning. Our findings reveal that finetuning smaller BERT-based embedding models (`gte-large-en-v1.5`, `roberta-large`) can outperform commercial embedding models (`text-embedding-3-small`, `text-embedding-3-large`) and top-performing public models (`SFR-Embedding-Mistral`, `gte-Qwen2-1.5B-instruct`, `gte-Qwen2-7B-instruct`). Notably, our benchmark evaluates model performance under temporal data splits and cross-report style transfer from GRI to ESRS, demonstrating the generalizability of the fine-tuned models.

In summary, our contributions are as follows:

- We create the ESG-Content Index Dataset (`ESG-CID`), a dataset leveraging disclosure content indices from ESG reports to facilitate research in the ESG domain and support the development of retrieval models for standardized ESG reporting.
- We benchmark state-of-the-art embedding models on `ESG-CID`, highlighting their limitations in the ESG retrieval task out of the box and demonstrating the benefits of domain-specific fine-tuning.
- We conduct detailed analyses of model performance under temporal splits and cross-report style transfer, offering insights into the challenges and solutions for automating ESG report generation, particularly in the context of the new ESRS standards.

Metric	Value
Unique Topics	11
Unique Sections	112
Total Datapoints	1230
Avg. Sections/Topic	10
Avg. Datapoints/Section	11
Sections with GRI Overlap	99
Sections without GRI Overlap	13
Sections GRI Overlap ratio	0.88
Datapoints with GRI Overlap	648
Datapoints without GRI Overlap	582
Datapoints GRI Overlap ratio	0.53

Table 1: ESRS Statistics and Overlap with GRI. The table presents counts for unique topics, sections, and datapoints, along with their averages in the ESRS guidelines from the official GRI-ESRS interoperability data[1]. Section overlap is counted if at least one datapoint in the section overlaps with a GRI datapoint

2 Related Work

The ESG domain has abundant public sustainability reports but lacks labeled data. Recent advancements in LLMs and PDF ingestion are bridging this gap. Vaghefi et al. (2023) demonstrates the potential of LLMs to transform the ESG domain with a Climate-change query specific chat interface called *ChatClimate* powered by LLMs. More recent studies, such as *ChatReport* (Ni et al., 2023) and *ClimRetrieve* (Schimanski et al., 2024), focus on Question Answering within this domain through RAG. These studies, however, are limited by their focus on a narrow set of queries and evaluations based on only 10-20 documents. In contrast, our approach covers a broad spectrum of ESG framework requirements and queries, supported by extensive training and evaluation data.

Distant supervision is a key concept in low-resource model training (Quirk and Poon, 2017; Qin et al., 2018). Polignano et al. (2022) first proposed using the GRI content index as distant supervision for ESG annotations, focusing on table identification via Optical Character Recognition and its role

[1] GRI-ESRS-Mapping.xlsx

(a) Industry Distribution

(b) Report Statistics

(c) Documents by Year

(d) Dataset Split

Figure 2: Dataset characteristics and challenges: (a) Industry distribution, showcasing the diversity of reporting sectors. (b) Report statistics (page count vs. average word count per chunk, sized by chunk count), highlighting the variability in report length and chunk size, which pose challenges for retrieval models. (c) and (d): Dataset splits (Train, Dev, Test GRI, Test ESRS), illustrating the chronological approach and the out-of-domain ESRS test set.

in sentiment analysis. Our work extends this by linking ESRS and GRI frameworks and advancing representation learning through RAG-based automated content index creation.

RAG is a framework that enhances text generation by retrieving relevant external information, improving accuracy and contextual relevance in NLP tasks (Lewis et al., 2020; Jiang et al., 2023). However, most works on ESG domain rely on proprietary embeddings such as OpenAI, which are difficult to adapt to specific needs and pose privacy risks for company data. We enhance retrieval by fine-tuning on ESG-specific content indexes, exploring whether cost-efficient fine-tuning with high-quality data and smaller models can match more resource-intensive methods. We employ fine-tuning techniques with models such as RoBERTa-large (Devlin et al., 2019; Liu et al., 2019) and Alibaba-NLP/gte-large-en-v1.5 (Li et al., 2023; Zhang et al., 2024), leveraging the Model Test Evaluation Benchmark (MTEB; Muennighoff et al. (2022)) to identify the best-performing models. Additionally, our study also evaluates ModernBert (Warner et al., 2024) to further understand the impact of domain-specific fine-tuning on retrieval.

3 ESG-CID

In line with our goal to enhance ESG-specific retrieval systems, we first collected a comprehensive set of sustainability and annual reports from companies across various industries and regions. Utilizing a combination of automated web crawling and manual collection techniques, we gathered over 10,000 reports from 2018 to 2023. The automated collection leveraged databases such as the now-decommissioned GRI database and the SRN database (Donau et al., 2023). After filtering out duplicates and non-English reports, we retained approximately 2,500 unique reports.

Out of these, around half adhered to the GRI standards, with a subset including the disclosure content index in a machine-readable format. We manually curated 73 GRI reports containing detailed content indices to form the primary dataset for our study. Additionally, we identified 11 reports from early adopters of the ESRS standards, which included ESRS content indices, enriching our dataset with cross-standard representations. The collected reports cover a diverse array of industries, predominantly from the financial, automotive, and manufacturing sectors (see Figure 2(a)).

3.1 Leveraging Content Indices for Weak Labeling

The disclosure content index serves as a structured bridge between the ESG standard requirements and the report content, providing an opportunity to create weakly labeled data without extensive manual annotation. Each content index lists the standard disclosure requirements (e.g., GRI or ESRS IDs and descriptions), along with references to the pages in the report where these disclosures are addressed.

As illustrated in Figure 2(b), the sustainability reports are significantly lengthy, averaging around 120 pages each, with the longest document exceeding 350 pages. Annotating such extensive documents is labor-intensive and impractical, especially when fine-grained annotations at the chunk or sentence level are considered. To address this challenge, we manually extracted only the content indices from the reports focusing only on these specific but crucial sections. Two experienced annotators, well-versed in ESG reporting and familiar with both GRI and ESRS standards, undertook this task. Their expertise ensured the accuracy and consistency of the extracted content indices.

Using the extracted content indices, we align the disclosure requirements with their corresponding page numbers in the reports. By automatically associating each standard query q (i.e., the disclosure requirement) with the relevant sections of the report indicated by the page numbers, we generate a set of query-document pairs. The query is a standard disclosure requirement, and the document is the corresponding page content addressing that requirement. Leveraging this inherent structure allows us to create a weakly labeled dataset suitable for training and evaluating retrieval models.

3.2 Creating Triplets for Embedding Models

To train and evaluate retrieval models in a contrastive learning framework, we construct triplets consisting of a query q, a positive (matched) chunk c^+, and a negative (unmatched) chunk c^-.

Positive Chunks We preprocess the PDF documents to segment them into manageable chunks (details in §C). The positive chunks c^+ are extracted from the pages referenced in the content index for each disclosure requirement. This ensures that c^+ contains information pertinent to the query q.

Negative Chunks For the negative samples c^-, we randomly sample chunks from the same report that are not associated with the given disclosure requirement. This assumes that these chunks are less relevant or irrelevant to the query, providing a contrastive signal for training.

3.3 Refining Labels with LLM Judgments

While the content indices provide page-level references, not all text within the referenced pages may directly address the disclosure requirement. To enhance the quality of our dataset, we employ Large Language Models (LLMs) as automated judges to assess the relevance of each chunk to the corresponding query.

We define a scoring function $s = \texttt{LLMScore}(q, c)$ that assigns a relevance score between 0 and 5 to each query-chunk pair. The LLM evaluates whether the chunk c sufficiently addresses the disclosure requirement q. By applying a relevance threshold (e.g., $s \geq 3$), we filter out positive chunks that are not sufficiently relevant, thus improving the quality of the triplets.

This refinement step ensures that our dataset contains high-quality, relevant query-document pairs, enhancing the effectiveness of retrieval models trained or evaluated on this data[2].

3.4 Dataset Splitting for Real-World Evaluation

To simulate real-world scenarios, particularly the temporal evolution of ESG standards and the adoption of new reporting requirements, we strategically split our dataset based on report release years and reporting standards.

Temporal Splitting The 73 GRI reports are ordered chronologically. We allocate the 10 most recent reports released after 2020, which adhere to the updated `GRI-NEW` standards, to form the test set ($\texttt{TEST} - \texttt{GRI}$). The next 5 most recent reports are designated as the development set for hyperparameter tuning. The remaining 58 reports, primarily following the older `GRI-OLD` standards, constitute the training set as shown in Fig 2(d). This split emulates a scenario where models trained on earlier data are evaluated on newer standards, testing their ability to generalize over time.

Cross-Standard Transfer The 11 ESRS reports form a separate test set ($\texttt{TEST} - \texttt{ESRS}$), allowing us to assess the models' performance on a different but related standard. This setup facilitates the

[2]Details on the LLM prompts and scoring criteria are provided in the §B

evaluation of cross-standard transferability and the models' adaptability to new reporting frameworks.

By organizing the dataset in this manner, we ensure that our evaluations reflect the challenges faced in real-world applications, such as adapting to evolving standards and handling reports from different time periods.

4 Experimental Setup

4.1 Embedding Models

We benchmark the retrieval performance of several state-of-the-art embedding models, including both LLMs and lightweight BERT-based models (< 1B Params). The LLM-based embeddings comprise open-source models such as `gte-Qwen2-1.5B-instruct`(Li et al., 2023), `gte-Qwen2-7B-instruct`(Li et al., 2023), and `SFR-Embedding-Mistral`(Rui Meng, 2024) , which are known for their strong capabilities in capturing complex language representations. We also include commercial models from OpenAI, namely `text-embedding-3-small` and `text-embedding-3-large`.

In addition to the LLMs, we evaluate lightweight BERT-based models suitable for deployment in resource-constrained environments. These include `roberta-large`(Liu et al., 2019), `ModernBERT-large`(Warner et al., 2024) and `gte-large-en-v1.5`(Li et al., 2023; Zhang et al., 2024) , which offer a balance between performance and computational efficiency. By comparing these models, we aim to understand the trade-offs between large-scale embeddings and more efficient alternatives in the ESG retrieval context.

4.2 Fine-tuning on `ESG-CID`

To enhance the domain-specific performance of the lightweight BERT-based models, we fine-tune them on the training split of our constructed dataset (`ESG-CID`). We utilize the standard Multiple Negatives Ranking Loss (Reimers and Gurevych, 2019) for contrastive learning using triplets consisting of a query, a positive chunk, and a negative chunk ((q, c^+, c^-)). Each query is associated with one relevant positive chunk and one irrelevant negative chunk, as detailed in Section 3.

The fine-tuning process spans five epochs. Further training details are provided in the Appendix. The fine-tuned models are referred to as `roberta-large`–`FT`, `ModernBERT-large`, and `gte-large-en-v1.5`–`FT`, respectively. We hypothesize that fine-tuning will imbue these models with ESG-specific knowledge, improving their retrieval capabilities on domain-specific queries.

4.3 Evaluation Metrics

We evaluate the models using standard retrieval metrics to assess their ability to rank relevant document chunks given a query. The metrics employed include Recall@20, which measures the proportion of relevant documents retrieved in the top 20 results; Mean Reciprocal Rank at 100 (MRR@100), indicating how early the first relevant document appears; Mean Average Precision at 100 (MAP@100), averaging precision scores at ranks where relevant documents are found; and Normalized Discounted Cumulative Gain at 100 (NDCG@100), emphasizing the ranking positions of relevant documents.

Performance is reported on both the GRI test split ($\texttt{TEST} - \texttt{GRI}$) and the ESRS test split ($\texttt{TEST} - \texttt{ESRS}$). It is noteworthy that the fine-tuned models were trained exclusively on the GRI training data and have not been exposed to any ESRS data, allowing us to evaluate their generalization capabilities across different ESG reporting standards.

4.4 Real-world Applicability: ESRS Content Indexing

Beyond standard retrieval metrics, we assess the practical utility of the models in constructing the ESRS content index within a company's report. According to ESRS, companies are required to provide structured disclosures in a tabular format. Our objective is to automate the extraction and indexing of relevant information from PDF reports according to each disclosure requirement.

In this task, given a document D and a set of ESRS disclosure queries $Q = \{q_1, q_2, \ldots, q_n\}$, we aim to map each query q_i to its corresponding page numbers in D. We experiment with reports from two companies—one in the automotive industry and one in agriculture—to capture diversity in reporting styles. We report the precision, recall and F1 of these mappings.

The models are employed within a Retrieval-Augmented Generation (RAG) framework. Each report D is segmented into chunks, and for each disclosure query q_i, the model retrieves the most relevant chunks from D. The retrieved chunks are then mapped back to their page numbers, effectively constructing the content index. Evaluation is based on the accuracy of these mappings, reflecting the models' effectiveness in automating the ESRS content indexing process.

Model	Size	TEST – GRI REC @20	MRR @100	MAP @100	NDCG @100	TEST – ESRS REC @20	MRR @100	MAP @100	NDCG @100
gte-Qwen2-1.5B-instruct	1.5B	0.72	0.38	0.34	0.48	0.44	0.22	0.19	0.29
gte-Qwen2-7B-instruct	7B	0.74	0.45	0.40	0.52	**0.47**	0.26	0.23	0.32
SFR-Embedding-Mistral	7B	0.66	0.36	0.31	0.45	0.42	0.23	0.19	0.28
text-embedding-3-small		0.70	0.40	0.36	0.49	0.43	0.21	0.18	0.28
text-embedding-3-large		0.74	0.47	0.41	0.53	0.46	0.27	0.23	0.32
roberta-large ❄	355M	0.28	0.11	0.08	0.21	0.19	0.08	0.06	0.15
ModernBERT-large ❄	396M	0.22	0.08	0.06	0.18	0.18	0.07	0.05	0.14
gte-large-en-v1.5 ❄	434M	0.69	0.40	0.36	0.48	0.43	0.21	0.18	0.28
roberta-large–FT	355M	0.74	0.43	0.38	0.51	0.40	0.21	0.17	0.27
ModernBERT-large–FT	396M	0.76	0.50	0.43	0.55	0.44	0.26	0.22	0.31
gte-large-en-v1.5–FT	434M	0.74	0.44	0.39	0.52	0.44	0.23	0.19	0.29
roberta-large–$FT_{LLMScore}$	355M	0.77	0.55	**0.50**	0.60	**0.47**	0.28	0.24	0.33
ModernBERT-large–$FT_{LLMScore}$	396M	**0.78**	**0.57**	**0.50**	**0.61**	0.46	0.28	0.24	0.33
gte-large-en-v1.5–$FT_{LLMScore}$	434M	**0.78**	0.56	**0.50**	**0.61**	**0.47**	**0.30**	**0.26**	**0.34**

Table 2: Overall effectiveness of the models on GRI Dev Set and ESRS Index. The best results are highlighted in boldface. For the GRI dataset, our fine-tuned models markedly outperform OpenAI, with our best performing fine-tuned model being better by up to 10% on all the ranking metrics, setting the state-of-the-art on this benchmark. For the ESRS dataset, our best performing fine-tuned model outperforms Open AI's text-embeddings-3-large by up to 2-3% on the ranking metrics. The low baseline underscores the significant challenge in ESRS disclosure retrieval.

5 Results and Analysis

5.1 Benchmarking Pre-trained Embedding Models

Table 2 presents the retrieval performance of various state-of-the-art embedding models on the GRI and ESRS test sets.

Firstly, we observe that most of the LLM-based embedding models demonstrate strong performance out of the box. For instance, the 1.5B parameter `gte-Qwen2-1.5B-instruct` embedding model achieves a Recall@20 of 0.72 without any domain-specific fine-tuning. Additionally, the open-source model `gte-Qwen2-7B-instruct` performs comparably to the commercial model `text-embedding-3-large`, highlighting the competitiveness of open-source solutions.

Secondly, LLM-based embedding models (listed in the first section of the table) significantly outperform the BERT-based embedding models (listed in the second section). This difference is attributed to the higher representational power and larger pre-training datasets of the LLM-based models, which enable better capture of semantic relationships in the ESG domain.

Thirdly, we note that the ESRS dataset presents a greater challenge compared to GRI. There is a substantial performance degradation across models when evaluated on ESRS, indicating that ESRS retrieval tasks are more difficult, possibly due to differences in standards or less overlapping training data.

5.2 Benchmarking Fine-tuned Embedding Models

We present the performance of our fine-tuned models in the last section of Table 2. While the original BERT-based models perform significantly worse than the LLM-based embeddings in their pre-trained state, fine-tuning on our dataset results in substantial performance improvements. After fine-tuning, the BERT-based models not only close the gap but, in most cases, outperform the larger LLM-based embeddings.

Specifically, for the GRI test set, `gte-large-en-v1.5`–`FT` achieves improvements of over 10 percentage points across all ranking metrics. Similarly, `roberta-large`–`FT` demonstrates consistent gains, outperforming the LLM-based models despite having fewer parameters. This showcases the effectiveness of fine-tuning on domain-specific data for enhancing model performance.

When evaluating the transfer performance to the ESRS test set, the fine-tuned models continue to perform significantly better than their pre-trained counterparts. Notably, the fine-tuned `gte-large-en-v1.5`–`FT` model outperforms the commercial baselines across all ranking metrics, despite not having been trained on any ESRS data. This suggests that fine-tuning on GRI data imparts transferable knowledge that generalizes to ESRS retrieval tasks.

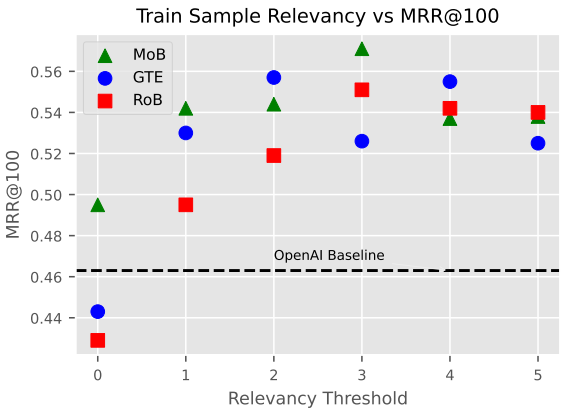

Figure 3: Relevancy Threshold vs MRR @100.

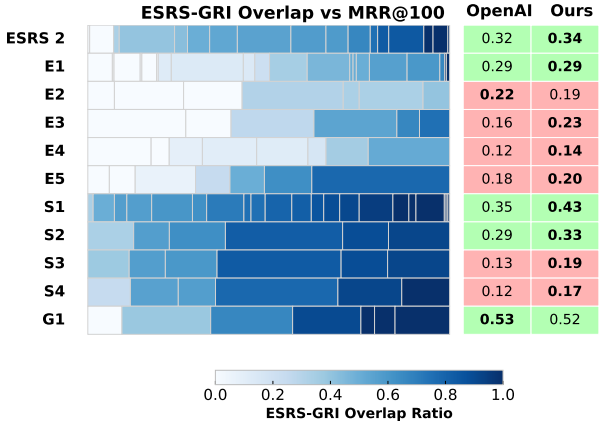

Figure 4: ESRS-GRI overlapping datapoints grouped by topics (top to bottom). Sections within each topic are ordered by their overlapping ratio (left to right). The table on the right displays ranking scores, using the MRR@100 metric, comparing OpenAI embeddings with those from our best-performing model. Scores from the better-performing model are bolded. Positive results (with MRR > 0.25) are highlighted in green, while negative results are highlighted in red.

5.3 Interplay between ESRS and GRI

To investigate the lower baseline scores observed on the ESRS test set, we performed a fine-grained analysis of the overlap between ESRS topics and GRI standards. The heatmap in Figure 4 illustrates the overlapping sections and is paired with the MRR@100 scores achieved by our best-performing models compared to the OpenAI baseline for each ESRS topic.

We identify topics `E2`, `E3`, `E4`, and `E5` as problematic due to insufficient training data, warranting further scrutiny. Similarly, topics `S3` and `S4`, despite having substantial training data, diverge from GRI mappings, indicating potential discrepancies in the ESRS-GRI correspondence. On the other hand, topics `ESRS 2`, `E1`, `S1`, `S2`, and `G1` yield strong performance, suggesting they are well-suited for automation. These topics show high overlap with GRI, reinforcing the potential to leverage existing GRI data for fine-tuning retrieval systems aimed at ESRS/CSRD-compliant reporting.

The problematic topics highlighted in red emphasize areas where additional data collection and methodological refinement are necessary to improve mapping accuracy. Future work may focus on enhancing the GRI-ESRS correspondence or incorporating additional standards into the training set to further boost ESRS performance.

5.4 Impact of `LLMScore` Filtering

To understand the contribution of the `LLMScore` filtering step, we conducted an ablation study on the GRI development set. Table 2 also compares the performance of the finetuned model with and without `LLMScore` filtering. Removing the `LLMScore` filtering step (i.e., using *all* triplets generated from the content index, regardless of the LLM's assessment) leads to a statistically significant drop in performance. This confirms that the LLM filtering helps to remove noise and improve the quality of the training data, leading to a more effective retrieval model.

5.5 Sensitivity of LLMScore Threshold

To determine the optimal threshold for filtering triplets using the `LLMScore`, we experimented with different threshold values on the DEV set. Figure 3 shows the MRR@100 performance of the finetuned models with thresholds ranging from 0 to 5. A threshold ≥ 1, which means keeping all triplets that have any positive relevance score from the LLM, provides better performance than doing no filtering. As the threshold increases, performance steadily increases indicating that having better quality samples improves retrieval performance. However, discarding triplets after a certain point (threshold ≥ 4) results in a less effective model. This shows for the LLM-provided grading signal, there exists an optimal threshold to maximize the utilization of the weakly supervised data.

5.6 ESRS Content Indexing

Table 3 presents the results of ESRS content indexing, comparing the performance of our finetuned `gte-large-en-v1.5-FT` model with the OpenAI embeddings. We observe that `gte-large-`

Company	Model	Prec	Rec	F1
Auto	text-embedding-3-large	0.30	0.41	0.34
	gte-large-en-v1.5 ❄	0.30	0.38	0.34
	gte-large-en-v1.5–FT	**0.36**	0.35	0.35
	gte-large-en-v1.5–FT$_{\texttt{LLMScore}}$	0.32	**0.49**	**0.39**
Agri	text-embedding-3-large	0.53	**0.50**	**0.52**
	gte-large-en-v1.5 ❄	0.55	0.49	**0.52**
	gte-large-en-v1.5–FT	**0.56**	0.46	0.50
	gte-large-en-v1.5–FT$_{\texttt{LLMScore}}$	0.54	0.44	0.48

Table 3: Comparison of GTE and OpenAI models for content index generation on an Automotive (Auto) and an Agricultural (Agri) companies.

en-v1.5–FT$_{\texttt{LLMScore}}$outperforms the OpenAI embeddings for the automotive company, whereas the OpenAI model performs better for the agriculture company. This discrepancy is likely due to the availability of fine-tuning data. Our training set contains abundant data from the automotive industry, which benefits the fine-tuned model. In contrast, the agricultural sector has limited representation in our training data, potentially disadvantaging the fine-tuned model compared to the more general-purpose OpenAI embeddings. Interestingly, LLMScore hurts the precision of the RAG system indicating that the models trained with LLM filtering confuse the RAG system by retrieving relevant looking false positives. Future work can refine the RAG through prompt tuning.

6 Conclusion

This paper addresses the critical need for scalable ESG information retrieval by leveraging disclosure content indices to align GRI and ESRS frameworks. Despite the abundance of publicly available sustainability reports, creating structured datasets has been challenging due to the labor-intensive nature of manual annotation. By using content indices as a source of weak supervision, we developed a novel benchmark for ESG retrieval finetuning and showed our models that outperform strong baselines, such as OpenAI.

Our results demonstrate that GRI indices can effectively bootstrap models for ESRS compliance, achieving moderate transferability despite limited ESRS-specific data. The LLMScore filtering process further enhanced training data quality, enabling our models to generalize across evolving ESG standards. These findings highlight the practical benefits of structured indices in automating ESG reporting and compliance tasks.

By harmonizing the GRI and ESRS frameworks, this research establishes a robust foundation for future inquiries into standard-agnostic capabilities, adaptability across regulatory frameworks, and holistic ESG reporting solutions. Our methodology significantly advances the field of ESG data retrieval, while also forging new paths for the creation of domain-specific LLMs tailored to meet the dynamic demands of sustainability regulations.

Limitations

- **Data Quality and Heterogeneity**: The reliance on content indices can introduce errors or omissions into the training dataset. The variability in reporting styles across industries complicates model generalization.
- **Transferability Across Standards**: Limited ESRS data may hinder the robustness of transfer learning from GRI to ESRS, potentially requiring frequent model updates as standards evolve.
- **LLMScore and Filtering Challenges**: The sensitivity of model performance to the LLMScore threshold indicates potential instability. Filtering might introduce false positives, impacting model precision.
- **Industry and Temporal Bias**: The dataset may be skewed towards certain industries, affecting model performance across different sectors. Temporal splits might not account for future changes in reporting practices.

Ethics Statement

In alignment with the ACL 2025 guidelines, we highlight the ethical aspects related to the participation of annotators in research activities. We are committed to ensuring that our approach to data annotation is humane, respectful, and inclusive, as this not only enhances the quality of the datasets but also respects and preserves the dignity and rights of all participants.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Charlotte-Louise Donau, Fikir Worku Edossa, Joachim Gassen, Gaia Melloni, Inga Meringdal, Bianca Minuth, Arianna Piscella, Paul Pronobis, and Victor Wagner. 2023. SRN Document Database. Accessed: 2023.

Marcelo Gutierrez-Bustamante and Leonardo Espinosa-Leal. 2022. Natural language processing methods for scoring sustainability reports—a study of nordic listed companies. *Sustainability*, 14(15):9165.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, et al. 2023. Chatreport: Democratizing sustainability disclosure analysis through llm-based tools. *arXiv preprint arXiv:2307.15770*.

Marco Polignano, Nicola Bellantuono, Francesco Paolo Lagrasta, Sergio Caputo, Pierpaolo Pontrandolfo, and Giovanni Semeraro. 2022. An NLP approach for the analysis of global reporting initiative indexes from corporate sustainability reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 1–8, Marseille, France. European Language Resources Association.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147, Melbourne, Australia. Association for Computational Linguistics.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Shafiq Rayhan Joty Caiming Xiong Yingbo Zhou Semih Yavuz Rui Meng, Ye Liu. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Tobias Schimanski, Jingwei Ni, Roberto Spacey, Nicola Ranger, and Markus Leippold. 2024. Climretrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. *Preprint*, arXiv:2406.09818.

Aida Usmanova and Ricardo Usbeck. 2024. Structuring sustainability reports for environmental standards with LLMs guided by ontology. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 168–177, Bangkok, Thailand. Association for Computational Linguistics.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, et al. 2023. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4(1):480.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*.

Yuchen Zhou and Alexander Perzylo. 2023. Ontosustain: Towards an ontology for corporate sustainability reporting. In *International Semantic Web Conference (ISWC)*.

A Hyperparameter settings

This section provides detailed information on the hyperparameter settings and training procedures used for fine-tuning the retrieval models (RoBERTa-large and GTE-large).

A.1 Hyperparameter Optimization

We used a combination of prior work, best practices for transformer fine-tuning, and empirical evaluation on a small validation set (carved out from the training set) to select the hyperparameters. Specifically, we held out five documents from the training set to form a validation set. This validation set was used solely for checkpoint selection and is distinct from the development set used for model evaluation. The primary metric for checkpoint selection was ‘dev_cosine_accuracy‘, defined below.

A.2 Training Arguments

Table 4 summarizes the key hyperparameters used for training. These settings were largely consistent across both RoBERTa-large and GTE-large, with the primary difference being the batch size due to GPU memory constraints.

Hyperparameter	RoBERTa-large	GTE-large
Training Epochs	5	5
Train Batch Size	32	8
Eval Batch Size	32	8
Warmup Ratio	0.05	0.05
FP16	False	False
BF16	False	False
Batch Sampler	No Duplicates	No Duplicates
Eval Steps	50	50
Save Steps	50	50
Save Total Limit	5	5
Logging Steps	20	20
Learning Rate	5e-5	5e-5
Load Best Model	True	True
Weight Decay	0.01	0.01
Metric for Best Model	‘cosine accuracy‘	‘cosine accuracy‘
DDP Find Unused Params	False	False

Table 4: Hyperparameter settings for fine-tuning RoBERTa-large and GTE-large.

We use saving and evaluation strategy based on the number of steps we take.

We used the ‘SentenceTransformerTrainingArguments‘ class from the ‘sentence-transformers‘ library to manage the training process. The key parameters are as follows:

- ‘output_dir‘: The directory where the trained models and checkpoints are saved. - ‘overwrite_output_dir‘: If ‘True‘, overwrites the contents of the output directory. - ‘num_train_epochs‘: The number of training epochs. We chose 5 epochs based on preliminary experiments, observing that performance plateaued after this point. - ‘per_device_train_batch_size‘: The batch size per GPU during training. We used a batch size of 32 for RoBERTa-large and 8 for GTE-large due to GPU memory limitations. - ‘per_device_eval_batch_size‘: The batch size per GPU during evaluation. - ‘warmup_ratio‘: The proportion of training steps used for a linear warmup of the learning rate. - ‘fp16‘ and ‘bf16‘: These were set to false due to hardware constraints. - ‘batch_sampler‘: We used the ‘NO_DUPLICATES‘ batch sampler, which ensures no duplicate examples within a batch. - ‘eval_strategy‘ and ‘eval_steps‘: Evaluation was performed every 50 training steps. - ‘save_strategy‘ and ‘save_steps‘: Model checkpoints were saved every 50 training steps. - ‘save_total_limit‘: Limited to 5 checkpoints to conserve disk space. - ‘logging_steps‘: Training statistics were logged every 20 steps. - ‘learning_rate‘: The initial learning rate for the AdamW optimizer was set to 5e-5. - ‘load_best_model_at_end‘: If ‘True‘, loads the model checkpoint with the best performance on the validation set at the end of training. - ‘weight_decay‘: The weight decay parameter for the AdamW optimizer. - ‘metric_for_best_model‘: The metric used for best model checkpoint selection was ‘eval_gri-chunk-dev_cosine_accuracy‘. - ‘ddp_find_unused_parameters‘: Set to ‘False‘ since distributed data parallel (DDP) training was not used.

A.3 Loss Function and Evaluation

The loss function used was ‘MultipleNegativesRankingLoss‘ from the ‘sentence-transformers‘ library. This loss function is designed for contrastive learning, ensuring that similar pairs (query and positive chunk) have higher similarity scores than dissimilar pairs (query and negative chunk). Each batch considered all other examples as negatives.

For development set evaluation, we used the ‘TripletEvaluator‘ from ‘sentence-transformers‘. The ‘TripletEvaluator‘ takes three lists as input:

- ‘anchors‘: A list of query examples. - ‘positives‘: A list of relevant chunks. - ‘negatives‘: A list of irrelevant chunks.

The evaluator computes the cosine similarity between anchor-positive and anchor-negative embeddings and calculates the ‘cosine_accuracy‘ metric.

A.4 Cosine Accuracy Metric

The ‘eval_gri-chunk-dev_cosine_accuracy‘ metric is calculated as follows:

1. Compute the cosine similarity between the query embedding and the positive chunk embedding: ‘sim_pos = cosine_similarity(M(q), M(c+))‘. 2. Compute the cosine similarity between the query embedding and the negative chunk embedding: ‘sim_neg = cosine_similarity(M(q), M(c-))‘. 3. Count the number of triplets where ‘sim_pos > sim_neg‘. 4. Compute ‘cosine_accuracy‘ as the percentage of triplets where the positive chunk has a higher cosine similarity to the query than the negative chunk.

This metric reflects the model’s ability to rank relevant chunks higher than irrelevant chunks.

A.5 Training Procedure

The models were trained using ‘MultipleNegativesRankingLoss‘, which is well-suited for contrastive training. Triplets of (query, positive chunk, negative chunk) were constructed, ensuring each query had one associated positive and one negative chunk. No significant overfitting was observed during the five training epochs.

B `LLMScore`Prompt Details

Below is the prompt used for ‘LLMScore‘, which leverages a Large Language Model (LLM) to assess the relevance of a text chunk to a given query, both extracted from an ESG report. The LLM is instructed to provide a numerical score on a scale of 0 to 5, reflecting the degree of relevance. See Figure 5 for further details.

`LLMScore` Prompt

Given the following `[query]`, and a `[text chunk]` from an ESG report, please rate the relevancy of the chunk to the disclosure on a scale of 0-5, in terms of being able to provide evidence for the disclosure. Provide higher rating if the chunk has enough evidence to answer the query.

- The output should be a single number between 0 and 5. 0 means not relevant at all, 5 means highly relevant.
- The output should be an integer

```
[query]
{disclosure}
[text chunk]
{chunk}
```
Relevancy Score (1-5): `<YOUR ANSWER HERE>`

Figure 5: Prompt for `LLMScore`

C PDF Preprocessing

For the ingestion of long sustainability PDF documents, we adopt the popular `PyMUPdfLoader` library with scalability in mind. After extracting the text from each page of the report we perform the following steps:

1. **Newline Removal:** Remove newline characters to produce continuous text.
2. **Chunking:** Partition the text on a pagewise basis into segments of 2048 characters.
3. **Overlap:** Apply an overlap of 512 characters between contiguous chunks to preserve context.

Formally, for a given PDF document $d \in \mathcal{D}$, the loader produces a set of text chunks:

$$\mathcal{C}(d) = \{c_1, c_2, \ldots, c_n\},$$

where each chunk c_i is a sequence of 2048 characters (with a 512-character overlap with c_i and c_{i+1}). These chunks serve as the basic units for further processing in our pipeline.

D Dataset Example

In this section, we provide examples of the GRI index and the ESRS index from the HYUNDAI 2024 sustainability report. This communicates the complexity of the existing pdf data and why generating an ESRS report from the the GRI format report is challenging. Additionally, once relevent ESRS index and GRI index are identified; collating related content is non-trivial.

Figure 6 shows details of the ESRS2, ESRS E1, and ESRS E2 index. Whereas Figure 7, Figure 8 and Figure 9 shows all the caption index for the GRI. We can also see how caption index in ESRS 2 have good overlap with caption index in GRI, whereas E1 has less overlap and E2 has least. This is also inline with our findings in the figure 4.

We have also included a few content examples from the Hyundai 2025 sustainability report to showcase how ESRS index differ from GRI index for the same content.

ESRS (European Sustainability Reporting Standards)

ESRS 2. General Disclosures

Indicator No.	Title	Page
ESRS 2 BP-1	General basis for preparation of the sustainability statements	124
ESRS 2 BP-2	Disclosures in relation to specific circumstances	28, 36, 42, 43, 97, 98, 100, 117-122
ESRS 2 GOV-1	The role of the administrative, management and supervisory bodies	9, 21, 81-85
ESRS 2 GOV-2	Information provided to and sustainability matters addressed by the undertaking's administrative, management and supervisory bodies	82, 85
ESRS 2 GOV-3	Integration of sustainability-related performance in incentive schemes	9, 17, 20, 37, 59
ESRS 2 GOV-4	Statement on sustainability due diligence	50-53, 67-69
ESRS 2 GOV-5	Risk management and internal controls over sustainability reporting[1]	-
ESRS 2 SBM-1	Market position, strategy, business model(s) and value chain	6-7, 25-26
ESRS 2 SBM-2	Interests and views of stakeholders	11-13
ESRS 2 SBM-3	Material impacts, risks and opportunities and their interaction with strategy and business model(s)	15-17
ESRS 2 IRO-1	Description of the processes to identify and assess material impacts, risks and opportunities	14
ESRS 2 IRO-2	Disclosure Requirements in ESRS covered by the undertaking's sustainability statements	110-112

[1] We have been operating an IT system-based "ESG platform" since 2022 to secure ESG data collection-inspection-disclosure efficiency and credibility of all business sites in Korea and abroad

ESRS E1. Climate Change

Indicator No.	Title	Page
ESRS E1-1	Transition plan for climate change mitigation	32
ESRS E1-2	Policies related to climate change mitigation and adaptation	23-32
ESRS E1-3	Actions and resources in relation to climate change policies	32, 37
ESRS E1-4	Targets related to climate change mitigation and adaptation	24-26, 30-32, 38
ESRS E1-5	Energy consumption and mix	98
ESRS E1-6	Gross Scopes 1, 2, 3 and Total GHG emissions	36, 98
ESRS E1-7	GHG removals and GHG mitigation projects financed through carbon credits	16, 31
	Avoided emissions of products and services	15, 27
ESRS E1-8	Internal carbon pricing[1]	-
ESRS E1-9	Potential financial effects from material physical and transition risks and potential climate-related opportunities	22, 33-35

[1] Internal carbon price is used for improving energy efficiency, implementing low-carbon strategies and investments, and discovering and harnessing low carbon business opportunities

ESRS E2. Pollution

Indicator No.	Title	Page
ESRS E2-1	Policies related to pollution	19, 43
ESRS E2-2	Actions and resources related to pollution	20, 43
ESRS E2-3	Targets related to pollution	44
ESRS E2-4	Pollution of air, water and soil	100
ESRS E2-5	Substances of concern and substances of very high concern	44
ESRS E2-6	Potential financial effects from pollution-related impacts, risks and opportunities	20

Figure 6: European Sustainability Reporting Standards (ESRS) data example from Hyundai 2025 report.

GRI Index

Universal Standards

GRI Standards No.	GRI Standards Title	Page	Note
2-1	Organizational details	124	
2-2	Entities included in the organization's sustainability reporting	-	p.464-468 of Business Report
2-3	Reporting period, frequency and contact point	124	
2-4	Restatements of information	28, 36, 42, 43, 97, 98, 100	
2-5	External assurance	117-123	
2-6	Activities, value chain and other business relationships	4, 5, 69	
2-7	Employees	101-103	
2-8	Workers who are not employees[1]	-	
2-9	Governance structure and composition	81-85	
2-10	Nomination and selection of the highest governance body	81	
2-11	Chair of the highest governance body	81	
2-12	Role of the highest governance body in overseeing the management of impacts	9, 21, 83, 85	
2-13	Delegation of responsibility for managing impacts	9, 21, 85	
2-14	Role of the highest governance body in sustainability reporting	85	
2-15	Conflicts of interest	81, 84, 87	
2-16	Communication of material issues	82, 85	
2-17	Collective knowledge of the highest governance body	83	
2-18	Evaluation of the performance of the highest governance body	83	
2-19	Remuneration policies	84	
2-20	Process to determine remuneration	84	
2-21	Annual total compensation ratio	84	
2-22	Statement on sustainable development strategy	3	
2-23	Policy commitments	19, 46, 50-51, 66, 88-89	
2-24	Embedding policy commitments	19, 46, 50-51, 66-69, 88-89	
2-25	Processes to remediate negative impacts	20, 53-54, 59	
2-26	Mechanisms for seeking advice and raising concerns	13, 54, 88-89	
2-27	Compliance with laws and regulations	105	

[1] Reason for non-disclosure: Confidentiality. We manage information on workers who are not employees but it is difficult to disclose information on workers who are not Hyundai employees due to company regulations..

GRI Standards No.	GRI Standards Title	Page	Note
2-28	Membership associations	104	
2-29	Approach to stakeholder engagement	12-13	
2-30	Collective bargaining agreements	57, 102	
3-1	Process to determine material topics	14	
3-2	List of material issues	15-17	
3-3	Management of material issues	15-17, 21-41, 57-61, 66-69, 71-73	

Topic Specific Standards - Economic

GRI Standards No.	GRI Standards Title	Page	Note
201-1	Direct economic value generated and distributed	98	
201-2	Financial implications and other risks and opportunities due to climate change	22-36	
201-3	Defined benefit plan obligations and other retirement plans	62	
201-4	Financial assistance received from government	98	
202-1	Ratios of standard entry level wage by gender compared to local minimum wage	103	
202-2	Proportion of senior management hired from the local community	101	
203-1	Infrastructure investments and services supported	104	
203-2	Significant indirect economic impacts	104	
205-1	Operations assessed for risks related to corruption	88-89	
205-2	Communication and training about anti-corruption policies and procedures	88-89	
205-3	Confirmed incidents of corruption and actions taken	88-89	
206-1	Legal actions for anti-competitive behavior, anti-trust, and monopoly practices	88	
207-1	Approach to tax	94	
207-2	Tax governance, control, and risk management	94	

Figure 7: GRI data example 1/3 from Hyundai 2025 report.

GRI Index

Topic Specific Standards - Environmental

GRI Standards		Page	Note
No.	Title		
301-1	Materials used by weight or volume	42, 98	
301-2	Recycled input materials used	42, 98	
301-3	Reclaimed products and their packaging materials	42	
302-1	Energy consumption within the organization	98	
302-2	Energy consumption outside of the organization	36	
302-3	Energy Intensity	98	
302-4	Reduction of energy consumption	23-24	
303-1	Interactions with water as a shared resource	42-43, 99	
303-2	Management of impacts related to wastewater	43, 100	
303-3	Water withdrawal	99	
303-4	Water discharge	99	
303-5	Water consumption	20, 42, 99	
304-1	Operational sites owned, leased, managed in, or adjacent to, protected areas and areas of high biodiversity value outside protected areas	46-48	
304-2	Significant impacts of activities, products and services on biodiversity	46-48	
304-3	Habitats protected or restored	46-48	
304-4	IUCN Red List species and national conservation list species with habitats in areas affected by operations	48	

GRI Standards		Page	Note
No.	Title		
305-1	Direct (Scope 1) GHG emissions	36, 98	
305-2	Energy indirect (Scope 2) GHG emissions	36, 98	
305-3	Other indirect (Scope 3) GHG emissions	36, 98	
305-4	GHG emissions intensity	36, 98	
305-5	Reduction of GHG emissions	23-32	
305-7	Nitrogen oxides (NOx), sulfur oxides (SOx), and other significant air emissions	100	
306-1	Waste generation and significant waste-related impacts	40-43	
306-2	Management of significant waste-related impacts	40-43	
306-3	Waste generated	100	
306-4	Waste diverted from disposal	43, 100	
306-5	Waste directed to disposal	100	
308-1	New suppliers that were screened using environmental criteria	67-68	
308-2	Negative environmental impacts in the supply chain and actions taken	69	

Figure 8: GRI data example 2/3 from Hyundai 2025 report.

GRI Index

Topic Specific Standards - Social

GRI Standards		Page	Note
No.	Title		
401-1	New employee hires and employee turnover	103	
401-2	Benefits provided to full-time employees that are not provided to temporary or part-time employees	56-57, 62	
401-3	Parental leave	62, 103	
403-1	Occupational health and safety management system	58	
403-2	Hazard identification, risk assessment, and incident investigation	58-59	
403-3	Occupational health services	59, 62	
403-4	Worker participation, consultation, and communication on occupational health and safety	58-59	
403-5	Worker training on occupational health and safety	58-61	
403-6	Promotion of worker health	62	
403-7	Prevention and mitigation of occupational health and safety impacts directly linked by business relationships	59-61	
403-8	Workers covered by an occupational health and safety management system	58-59	
403-9	Work-related injuries	58-59, 105	
403-10	Work-related ill health	58-59, 105	
404-1	Average hours of training per year per employee	102	
404-2	Programs for upgrading employee skills and transition assistance programs	55-56	
404-3	Percentage of employees receiving regular performance and career development reviews	54	

GRI Standards		Page	Note
No.	Title		
405-1	Diversity of governance bodies and employees	81, 101-102	
405-2	Ratio of basic salary and remuneration of women to men	103	
406-1	Incidents of discrimination and corrective actions taken	54, 88	
407-1	Operations and suppliers in which the right to freedom of association and collective bargaining may be at risk	52, 69	
408-1	Operations and suppliers at significant risk for incidents of child labor	52, 69	
409-1	Operations and suppliers at significant risk for incidents of forced or compulsory labor	52, 69	
411-1	Incidents of violations involving rights of indigenous peoples	-	No incidents of violations occurred
413-1	Operations with local community engagement, impact assessments, and development programs	12, 46-48, 76-79, 104	
414-2	Negative social impacts in the supply chain and actions taken	69	
415-1	Political contributions	104	No political contributions made
416-1	Assessment of the health and safety impacts of product and service categories	73	
416-2	Incidents of non-compliance concerning the health and safety impacts of products and services	72, 105	
417-1	Requirements for product and service information and labeling	75	
417-2	Incidents of non-compliance concerning product and service information and labeling	105	No incidents of violations occurred
417-3	Incidents of non-compliance concerning marketing communications	105	No incidents of violations occurred
418-1	Substantiated complaints concerning breaches of customer privacy and losses of customer data	105	

Figure 9: GRI data example 3/3 from Hyundai 2025 report.

Example 1

ESRS 2 GOV-4: Statement on sustainability due diligence — Page no: 50-53, 67-69
GRI 2-24: Embedding policy commitments — page no 19, 46, 50-51, 66-69, 88-89
Section name: Human Rights and Human Resources Management

Hyundai supports international standards and guidelines related to human rights and labor, and promotes human rights management across global supply. In collaboration with the relevant departments, we strive to make practical improvements, while also conducting annual due diligence across our business sites and suppliers to identify both potential and actual human rights risks, and implementing appropriate mitigation measures accordingly. Meanwhile, we have established a human resources management system that provides the highest level of value to employees. We recruit talented employees and invest in capacity building to create a culture of voluntary learning. We also have built a creative and performance-oriented organizational culture performance evaluation and fair compensation, operate customized welfare systems, and carry out activities aimed at improving the work environment and promoting diversity.

E Future Work

Our work opens promising avenues for advancing ESG information retrieval, both technically and practically. One key direction is improving the automation and validation of content index extraction from diverse PDF formats, which necessitates robust table detection and structure recognition (**?**), alongside accurate semantic role labeling and validation using LLMs to ensure alignment with report content. Improved OCR techniques are also crucial.

Another significant area is expanding to multi-document, multi-linguality and multi-source retrieval, an essential step for a holistic view of a company's ESG performance. This involves challenges in cross-document coreference resolution, information fusion where conflicting data must be reconciled, and temporal reasoning to monitor changes over time. This approach could leverage existing methodologies in multi-source information retrieval but applied specifically to ESG contexts.

Regarding semantic and reasoning capabilities, our focus should shift towards integrating deeper understanding through ESG-specific knowledge graphs, enabling numerical and logical reasoning over tabular data and figures, and adopting more contextualized retrieval strategies. This is where bridging standards mapping becomes crucial, enhancing existing open-source mappings like RSO (Zhou and Perzylo, 2023; Usmanova and Usbeck, 2024) by co-learning across different frameworks to enrich our understanding and mapping capabilities.

The robustness, adaptability, and multilinguality of our models are vital due to the evolving nature of ESG standards. Research should push towards continual learning, few-shot/zero-shot learning for new disclosures, and cross-lingual transfer learning, especially considering the multilingual demands of the CSRD in the EU (Gutierrez-Bustamante and Espinosa-Leal, 2022).

Finally, our vision for an end-to-end ESGLLM model aims at automating the entire ESG reporting and analysis cycle. This includes not only report generation and summarization but also risk assessment, anomaly detection, and consistency/gap analysis, thereby simplifying processes for companies, investors, and regulators. This holistic approach could potentially transform how ESG data is handled, analyzed, and reported.

Example 2

ESRS E2-2: Actions and resources related to pollution — page 20 and 43
GRI 303-5: Water consumption — page 20, 42, 99
MANAGEMENT OF ENVIRONMENTAL PERFORMANCE

Management of Environmental Goals Through our environmental management implementation system, we set mid- to long-term performance goals for environmental factors that have a considerable environmental impact due to business operations, such as carbon emissions. Mid- to long-term performance goals are set in consideration of business as usual (BAU) as well as external economic circumstances, government policy direction, and internal business strategies. To respond to climate change, we set the goal to achieve carbon neutrality by 2045 throughout the entire life cycle. To achieve the goal, we are implementing such strategic tasks as a strategy to transition to EVs, achieving RE100 at business sites, and reduction of supply chain carbon emissions. For quantitative improvements to environmental indexes, excluding carbon, we set improvement goals for water and wastes based on the direction of suppressing increases in water consumption and waste generation that are on the rise in connection with production that is increasing after COVID-19. Additionally, we manage pollutant emissions at each business site – air (dust, NOx, SOx, THC) and water (TOC, TP, BOD, SS) – to stricter standards than the legal requirements, thereby strengthening our environmental pollutant management. We have also set an upper limit of 5% for the three-year average for pollutant emissions and established specific emission targets for each busines

Example 3

GRI 305-1: Direct (Scope 1) GHG emissions, page 36 and page 98
ESRS E1-7: Gross Scopes 1, 2, 3 and Total GHG emissions on page 36 and page 98
Climate related metrics
Scope 1 and Scope 2 Emission [1)]

classification	2021	2022	2023
Scope 1	724,013	719,949[2)]	696,590
Scope 2 (location-based)[3)]	1,853,813	1,831,531	1,831,531
Scope 2 (market-based)	1,660,058	1,684,120	1,579,161
Scope 1 + Scope 2[4)]	2,384,071	2,404,069	2,275,751
Scope 1 + Scope 2 Emission intensity (GHGs emissions per vehicle produced)	0.616	0.601	0.531

Example 4

ESRS E4-5: Impact metrics related to biodiversity and ecosystems change — page no 46-48
GRI 304-2: Significant impacts of activities, products and services on biodiversity — page no 46-48
Protection of Biodiversity

Biodiversity is essential for life on Earth, allowing humans, plants, and animals to live in harmony with nature. Recognizing that biodiversity has a significant impact on natural capital—including human food safety, health, air and water quality, and raw material supply—Hyundai strives to assess its impacts on, and risks to, biodiversity and to ameliorate any negative impacts based on this assessment. Furthermore, under the company-wide ”Colorful Life” campaign, we aim to prevent further loss of biodiversity and turn it into a net gain by implementing various projects, such as protecting endangered species and preserving natural habitants within the communities near our sites and regenerating land and marine ecosystems while taking into account their natural characteristics.

Example 5

ESRS E1-1: Transition plan for climate change mitigation - page no 32
GRI-305-5: Reduction of GHG emissions - page no 23-32
Plans to Achieve Climate-Related Targets (Carbon Neutrality Targets)

Reducing Our Carbon Emissions at Work Hyundai is a supporter for the Paris Agreement and recognizes its corporate role and responsibility to reduce global GHG emissions. In this regard, we strive to achieve carbon neutrality at our business sites by 2045 by switching to renewable energy, improving the energy efficiency of production processes through the introduction of high-efficiency motors and inverters, and utilizing hydrogen energy. In the short term, in conjunction with the RE100 roadmap, we plan to promote the transition from electric energy used in the manufacturing process to renewable energy first. In the long term, our goal is to achieve carbon neutrality by 2045 by expanding the application of green hydrogen and the use of renewable energy in conjunction with the realization of a hydrogen society.