Unsupervised GRI-TCFD Alignment with LLM-Assisted Validation for Climate Disclosure and Greenwashing Risk Analysis

Seyed Alireza Mousavian Anaraki, Danilo Croce,
Roberta Costa, Luigi Tiburzi, Armando Calabrese and Roberto Basili
Department of Enterprise Engineering
University of Rome Tor Vergata, Via del Politecnico 1, 00133, Rome, Italy
seyedalireza.mousaviananaraki@students.uniroma2.eu
{croce, basili}@info.uniroma2.it
{roberta.costa, luigi.tiburzi}@uniroma2.it
calabrese@dii.uniroma2.it

Abstract

Climate-related corporate disclosures play a central role in sustainable finance and regulatory supervision, but remain difficult to analyze due to their length, unstructured format, and strategic language. While existing NLP approaches have been applied to ESG scoring and greenwashing detection, most operate at the document level and lack explicit alignment with formal reporting standards. We propose a scalable paragraph-level framework for aligning sustainability disclosures with the Global Reporting Initiative (GRI) indicators and the Task Force on Climate-related Financial Disclosures (TCFD) pillars. Our approach combines weak supervision, climate-focused GRI-TCFD mapping, embedding-based semantic similarity, and LLM validation for climate detection. In parallel, we introduce a paragraph-level greenwashing proxy based on commitment intensity, claim specificity, and sentiment polarity. This proxy complements regulatory alignment by capturing linguistic signals associated with potentially symbolic climate communication. The resulting augmented data are used to fine-tune ClimateBERT models in both single-task and multi-task settings. Experimental results show that weakly supervised dataset augmentation improves robustness and generalization compared to purely manual training, with further gains in the multi-task configuration. By integrating regulatory semantics, domain-adapted language models, and scalable annotation strategies, this study advances standard-aligned climate disclosure analysis and provides tools directly relevant to climate-related financial risk assessment.

Keywords: Sustainability reporting, GRI, TCFD, Greenwashing Risk Proxy, ClimateBERT

1. Introduction

Climate-related disclosure has become a central pillar of sustainable finance and regulatory oversight. Institutional investors, banks, and supervisory authorities increasingly rely on structured reporting frameworks to assess exposure to transition and physical risks. Among the most widely adopted standards, the Global Reporting Initiative (GRI) provides granular sustainability indicators, while the Task Force on Climate-related Financial Disclosures (TCFD) defines a financial-material framework organized around Governance, Strategy, Risk Management, and Metrics and Targets. These frameworks directly inform regulatory supervision, capital allocation, and climate risk modeling (Bingler et al., 2022).

At the same time, sustainability reporting has expanded dramatically in volume and complexity. Reports are typically long, unstructured PDF documents written in strategic corporate language, with climate-related information scattered across sections (Anaraki et al., 2025). This makes paragraph-level interpretation difficult. Moreover, a growing body of literature reports on selective disclosure and "*greenwashing*" practices, in which narrative emphasis may exceed substantive commitment (Janik and Ryszko, 2025). Empirical evidence shows that discursive Environmental, Social, and Governance (ESG) emphasis may not always translate into proportional strategic or operational changes (Bingler et al., 2024; Wood et al., 2025).

Recent advances in natural language processing (NLP) and large language models (LLMs) have enabled automated extraction, classification, and verification of sustainability-related information (Moodaley and Telukdarie, 2023a). Transformer-based and LLM-based approaches have been used to detect green practices in social media (Glazkova and Zakharova, 2025), assess ESG commitment in financial documents (Wood et al., 2025), identify green claims and greenwashing (Moodaley and Telukdarie, 2023b), and perform climate-specific classification through domain-adapted models such as ClimateBERT (Webersinke et al., 2022). These tools support scalable and systematic analysis of corporate disclosures across ESG dimensions (Zou et al., 2025; Kazakov et al., 2023), facilitating better alignment with reporting frameworks such as GRI (Ngee et al., 2024; Bronzini et al., 2024), the Sustainable Development Goals (SDGs) (Jakob et al., 2024;

Proceedings of the 2nd Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology) @ LREC 2026, pages 15–25

Li and Rockinger, 2024), and TCFD, particularly through domain-adapted BERT variants such as ClimateBERT for greenwashing detection (Bingler et al., 2022). Prior work has also explored weakly supervised and unsupervised methods for automatically aligning report content with multiple sustainability standards, particularly focusing on integrating GRI and SDG (Mousavian Anaraki et al., 2025a,b).

However, existing work largely focuses on document-level scoring, sentiment detection, or general climate relevance, rather than structured paragraph-level alignment with formal regulatory architectures.

We address this gap by proposing a scalable paragraph-level annotation framework that leverages the officially published alignment between GRI indicators and TCFD pillars to guide climate-related text identification. Rather than directly training a regulatory classifier, we use GRI-TCFD alignment as structured weak supervision to construct high-confidence paragraph-level climate annotations. Our approach combines weak supervision from GRI content indices, embedding-based semantic similarity, and LLM-assisted validation. MPNet (Song et al., 2020) embeddings rank paragraph-to-definition matches across candidate GRI-TCFD pairs, while GPT-OSS (Agarwal et al., 2025) filters noisy assignments through definition-grounded reasoning. This process yields a standard-informed climate detection dataset without requiring manual paragraph-level expert annotation. For instance, the following excerpt from a sustainability report:

> "*Our goal is to be a leader in ecologically sound production by looking at our own carbon footprint and aiming for the highest possible standards of ecological responsibility among our producers.*"

is automatically classified as climate-related and aligned with the corresponding GRI-TCFD categories:

- **GRI 305** (Emissions), specifically **GRI 305-5**: "*Reduction of GHG emissions*"
- **TCFD** (Governance): "*Disclose the organization's governance around climate-related risks and opportunities*"
- **Climate-related** (Emissions, Governance)

In parallel, we introduce a paragraph-level greenwashing risk proxy derived from commitment intensity, claim specificity, and sentiment polarity. Inspired by (Vinella et al., 2024), we define a greenwashing risk proxy based on three characteristic patterns commonly observed in misleading sustainability disclosures: (1) absence of explicit climate-related commitments and actions, (2) use of non-specific or vague language, and (3) overly optimistic or promotional sentiment. Each dimension is treated as an indicator, and their combination yields a greenwashing risk proxy, which is discretized into *Low-risk* (one active indicator), and *High-risk* (two or more active indicators) categories. Following this formulation, the paragraph is labeled as:

- Commitment: Inactive
- Non-Specific: Active
- Optimistic Sentiment: Active

resulting in a High Greenwashing Risk label. The proxy is computed only on validated climate-related paragraphs and captures linguistic patterns associated with potentially symbolic climate communication. Rather than replacing regulatory alignment, this proxy complements climate relevance by providing an additional supervisory signal grounded in discourse characteristics. The automatically constructed datasets are then used for benchmark augmentation.

We fine-tune ClimateBERT in single-task (climate relevance) and multi-task (climate relevance and greenwashing risk) settings to evaluate whether standard-informed weak supervision improves downstream performance. By coupling regulatory structure with scalable annotation and multi-task learning, our framework reduces reliance on costly manual labeling while maintaining semantic consistency with established reporting standards and improving empirical robustness in downstream classification.

This paper makes three main contributions:

1. We propose a scalable, weakly supervised framework for climate-related paragraph-level alignment of corporate disclosures with GRI and TCFD standards, combining semantic similarity and LLM-based validation;
2. We introduce a transparent, literature-grounded paragraph-level greenwashing risk proxy based on commitment, specificity, and sentiment attributes;
3. We show that automatically generated annotations improve climate detection and greenwashing risk classification through dataset augmentation and multitask learning.

To guide our empirical evaluation, we address the following research questions:

- **RQ1.** Does augmenting benchmark datasets with automatically constructed annotations improve downstream performance?
- **RQ2.** Does joint multi-task learning provide additional gains over separate single-task models?

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed paragraph-level GRI-TCFD alignment framework and the greenwashing risk proxy construction. Section 4 reports the experimental evaluation. Section 5 concludes the paper.

2. Related Work

Natural language processing has been widely applied to the analysis of corporate sustainability disclosures. Prior work employs sentence similarity, sentiment classification, and information extraction techniques to assess environmental risk and ESG communication patterns in unstructured reports (Kang and Kim, 2022; Polignano et al., 2022). These approaches enable large-scale analysis but typically operate at the document level or focus on isolated linguistic signals.

The emergence of LLMs has further expanded capabilities for extracting structured information and detecting misleading or greenwashing claims in sustainability reporting (Moodaley and Telukdarie, 2023a,b). Domain-adapted models such as ClimateBERT (Webersinke et al., 2022) incorporate climate-specific pretraining and have demonstrated strong performance across tasks including climate risk detection (Garrido-Merchán et al., 2026), cheap talk analysis (Bingler et al., 2024), and greenwashing identification (Vinella et al., 2024).

However, progress remains constrained by the scarcity of high-quality paragraph-level annotations. Recent efforts such as Climate-NLI (Yudanto et al., 2024) explore zero- and few-shot classification through natural language inference (Yin et al., 2019), yet performance is sensitive to semantic overlap and task ambiguity.

In parallel, weakly supervised and unsupervised approaches have been proposed to align sustainability disclosures with structured reporting standards, particularly for integrating GRI and SDG frameworks (Mousavian Anaraki et al., 2025a,b). Nevertheless, structured paragraph-level alignment with formal climate-related architectures such as the GRI-TCFD mapping remains underexplored.

Our work builds on these strands by combining structured weak supervision, embedding-based similarity modeling, and LLM-assisted validation to construct a paragraph-level, standard-informed climate dataset. Unlike prior studies that primarily address document-level scoring or standalone classification tasks, we explicitly leverage the official GRI-TCFD alignment to guide annotation and evaluate its impact through benchmark augmentation and multi-task fine-tuning.

3. Standard-Informed Paragraph-Level Annotation Framework

This section presents a scalable, weakly supervised pipeline for constructing paragraph-level annotated datasets for (i) climate relevance detection and (ii) greenwashing risk classification from corporate sustainability reports. The framework follows a sequential two-stage design, where the output of the first stage directly feeds into the second. In the first stage, we construct a climate detection dataset by identifying climate-related paragraphs through GRI-TCFD alignment, embedding-based semantic similarity, and LLM-assisted validation. The objective is to maximize annotation precision while avoiding manual expert labeling, leveraging structured weak supervision from reporting standards. In the second stage, we derive a greenwashing risk dataset by restricting the analysis to the validated climate-related subset and assigning commitment, specificity, and sentiment attributes via few-shot prompting, following (Bingler et al., 2024). These attributes are aggregated into a composite greenwashing risk proxy. By grounding greenwashing assessment in previously validated climate-related paragraphs, the framework ensures conceptual consistency between climate relevance and risk characterization. The automatically constructed datasets are subsequently used for benchmark augmentation and downstream evaluation (Section 4).

3.1. Automatic Climate Paragraph Detection via GRI-TCFD Alignment

Our approach is inspired by prior work on automatic alignment between GRI and SDG standards (Mousavian Anaraki et al., 2025b), which demonstrated the effectiveness of combining weak supervision, semantic similarity, and structured standards for large-scale sustainability annotation. Building upon this foundation, we extend the methodology to the climate reporting domain by integrating the TCFD framework and introducing LLM-assisted validation. The objective is to construct a high-quality paragraph-level dataset D, with $|D| = M$, derived from N different corporate disclosures (paragraphs) (with $M < N$), augmented with standardized climate reporting categories, without requiring manual expert annotation. An overview of the framework is shown in Figure 1. The proposed framework consists of six main stages: (i) paragraph extraction and preprocessing, (ii) weak GRI label initialization from content indices, (iii) climate-focused GRI–TCFD pairing, (iv) embedding-based similarity scoring, (v) indexed and non-indexed label disambiguation, and (vi) LLM-assisted validation.

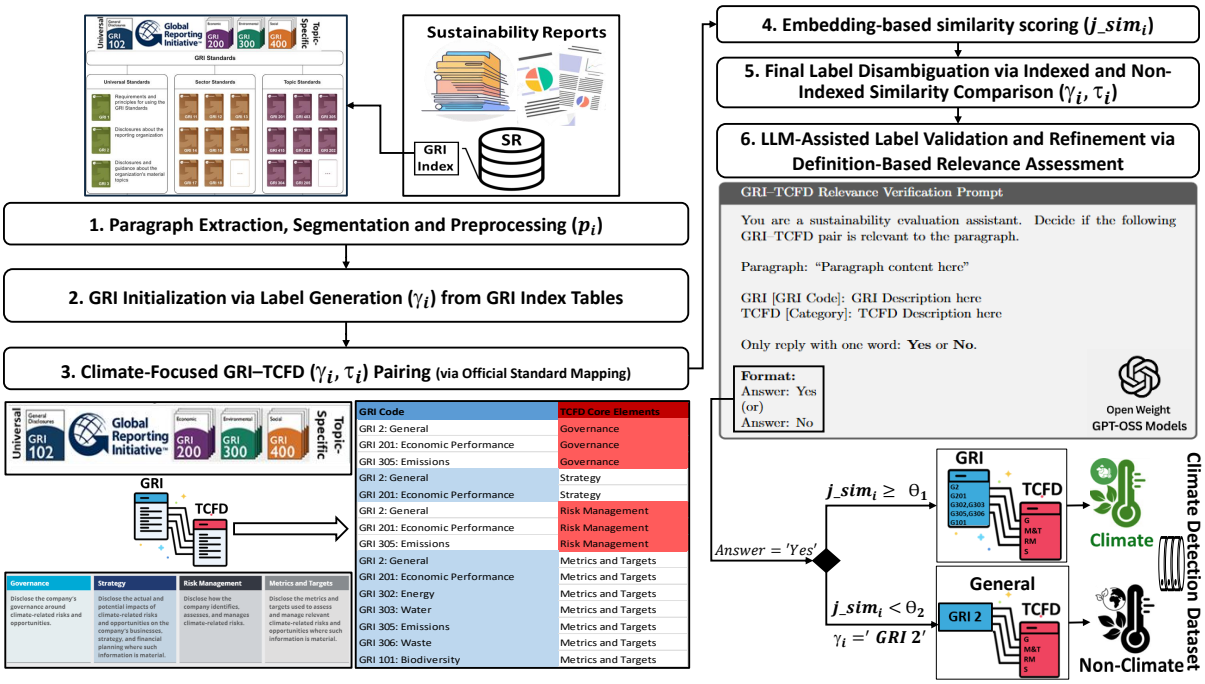

Figure 1: Overview of the proposed GRI–TCFD-based climate-related paragraph detection pipeline. The framework integrates weak supervision, semantic similarity modeling, and LLM-based validation.

The output is a structured climate detection dataset:

$$D = \{(p_i, \gamma_i, \tau_i, d(\gamma_i), d(\tau_i), llm_i, j\_sim_i, c_i)\}_{i=1}^{M},$$

where p_i is a paragraph, γ_i and τ_i its associated GRI and TCFD respective labels, $d(\gamma_i)$ and $d(\tau_i)$ their label descriptions, llm_i the *Yes/No* LLM validation output, j_sim_i the joint similarity score , c_i the final climate label, and M the number of samples after LLM validation (whereas $llm_i = Yes$).

Paragraph Extraction, Segmentation, and Preprocessing. Sustainability reports are extracted from PDF documents using layout-aware parsing with `PyMuPDF`. Headers, footers, and fragmented blocks are removed, with text blocks containing at least 20 words retained, as candidate paragraphs.

Weak GRI Annotation Initialization via Indexed and Non-Indexed Label Generation from the GRI Index Table. Most sustainability reports include a GRI content index mapping disclosure codes to page numbers. Although self-reported, this index provides valuable weak supervision.

For each paragraph p on page π, we define:

- **Indexed set**: `GRI` codes explicitly linked to π in the index.
- **Non-Indexed set**: all remaining `GRI` codes not mentioned in the index for π, but potentially relevant according to the semantic content.

This dual-set strategy mitigates incomplete coverage and potential strategic under-reporting.

Climate-Focused GRI-TCFD Pairing via Official Standard Mapping. To ensure conceptual consistency, we adopt the official GRI-TCFD alignment published by the GRI organization. This mapping ($\mathcal{M}$) restricts candidate labels to climate-relevant combinations. We focus on seven core GRI codes (GRI 2, 201, 302, 303, 305, 306, 101), yielding 15 valid GRI–TCFD pairs. These pairs define the candidate annotation space. For each paragraph p and each candidate GRI code γ, we generate triples (p, γ, τ) where $\tau \in \mathcal{M}(\gamma)$, such as:
`GRI 305:Emissions` $\rightarrow$ `TCFD:Governance`.

Embedding-based Similarity Scoring. To rank candidate label pairs (γ, τ), we compute semantic similarity between paragraphs and standard descriptions using the MPNet encoder (Song et al., 2020). Given the paragraph p and the texts connected to the GRI disclosure requirement and TCFD definition for γ and τ respectively, we encode them into fixed-dimensional embedding vectors. Let $\mathbf{e}_p$ denote the embedding of paragraph p. For each GRI code γ, let $R_\gamma = \{r_1, \ldots, r_{|R_\gamma|}\}$ represent the set of official disclosure texts associated with γ. Similarly, for each TCFD category τ, let $D_\tau = \{d_1, \ldots, d_{|D_\tau|}\}$ denote the set of textual definitions and guidance statements describing τ. For each triple (p, γ, τ), we compute a joint similarity score given by:

$$j\_sim_p = \max_{r \in R_\gamma, d \in D_\tau} \cos(\mathbf{e}_p, \mathbf{e}_r) \cdot \cos(\mathbf{e}_p, \mathbf{e}_d).$$

The multiplicative formulation enforces joint relevance to both reporting frameworks.

Final Label Disambiguation via Indexed and Non-Indexed Similarity Comparison. For each paragraph, we compare the highest-ranked indexed and non-indexed candidate triples. The candidate pair (γ, τ) with the higher similarity score j_sim_p is selected as the final label for each p, regardless of whether it comes from the indexed or non-indexed set. This conservative strategy prioritizes precision in the annotations.

LLM-based Validation. 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. To further improve annotation quality, embedding-based similarity is followed by LLM-assisted validation using GPT-OSS 20B (Agarwal et al., 2025). Each retained triple is evaluated through a structured prompt containing the paragraph and official definitions. In line with (Mousavian Anaraki et al., 2025a), the model outputs a binary relevance decision, approximating expert judgment and substantially reducing false positives. By filtering only on "*Yes*" LLM outputs and using Θ_1 as the minimum allowed joint similarity threshold, we assign final climate labels to paragraphs as climate-related. By filtering only on "*GRI 2 (General)*" with a maximum allowed threshold Θ_2, non-climate paragraphs can also be recognized. The resulting climate-related paragraphs constitute a reliable input for the later greenwashing risk analysis and evaluation described in the following subsection.

3.2. Greenwashing Risk Proxy Construction

Building on the climate-related annotations obtained in Section 3.1, we construct a composite greenwashing risk proxy by classifying commitment, specificity, and sentiment attributes using few-shot, in-context prompting. We emphasize that the proposed indicator represents a proxy for potential greenwashing risk based on textual characteristics, not a direct measure of deceptive corporate behavior.

To quantify the extent to which corporate climate-related disclosures may exhibit characteristics of greenwashing, we construct a composite greenwashing risk proxy based on linguistic and semantic attributes identified in prior literature (Bingler et al., 2024; Vinella et al., 2024). Bingler et al. (2024) introduces ClimateBERT-based downstream tasks to analyze firm-level climate communication. In particular, their methodology relies on fine-tuned language models to classify paragraphs according to climate relevance, sentiment, corporate commitments and actions, and linguistic specificity. These tasks are subsequently aggregated to form the Cheap Talk Index, which measures the proportion of non-specific climate-related commitments in corporate disclosures.

Building on this framework, we adopt the same core dimensions, *commitment*, *specificity*, and *sentiment*, as fundamental components for assessing the credibility of climate-related statements. Vinella et al. (2024) suggests that greenwashing is commonly associated with three main linguistic patterns: (1) the absence of explicit climate-related commitments, (2) the use of non-specific or vague language, and (3) the use of overly positive or optimistic sentiment without corresponding substantive actions. Following this literature, we operationalize greenwashing risk by combining indicators that capture these characteristics at the paragraph level. Figure 2 illustrates the second part of our methodology for greenwashing risk proxy construction after climate relevance identification from the previous part (3.1). The proposed methodology consists of three main stages : (i) attribute-based paragraph classification via prompting, (ii) automatic dataset extension, and (iii) construction of a composite greenwashing risk proxy GWR for downstream classification and benchmark augmentation.

Attribute-Based Classification via Prompting. For each climate-related paragraph p_i, we automatically infer three linguistic and semantic attributes: commitment Co_i, specificity Sp_i, and sentiment Se_i. These attributes correspond to key dimensions of climate communication defined in prior work (Bingler et al., 2024; Vinella et al., 2024). We design three task-specific prompts, each including two representative examples (2-shot), sampled from the corresponding ClimateBERT benchmarks, namely `climate commitments actions` (Bingler et al., 2023a), `climate specificity` (Bingler et al., 2023d), and `climate sentiment` (Bingler et al., 2023c). The prompts are constructed based on the formal task definitions provided in the benchmark documentation and are used to classify each paragraph as follows (Bingler et al., 2024):

- **Commitment:** "*A paragraph gets labeled as commitment-yes if it reports that the company undertook activities in this regard, if it reports that it will likely do so, or if the company sets targets in this paragraph.*"
- **Specificity:** "*A paragraph gets labeled as specific if a paragraph contains detailed performance information, details of actions, or tangible and verifiable targets.*"
- **Sentiment:** "*A paragraph is labeled as an opportunity if it mainly discusses business opportunities or the positive impacts of mitigating or adapting to climate change. It is considered*

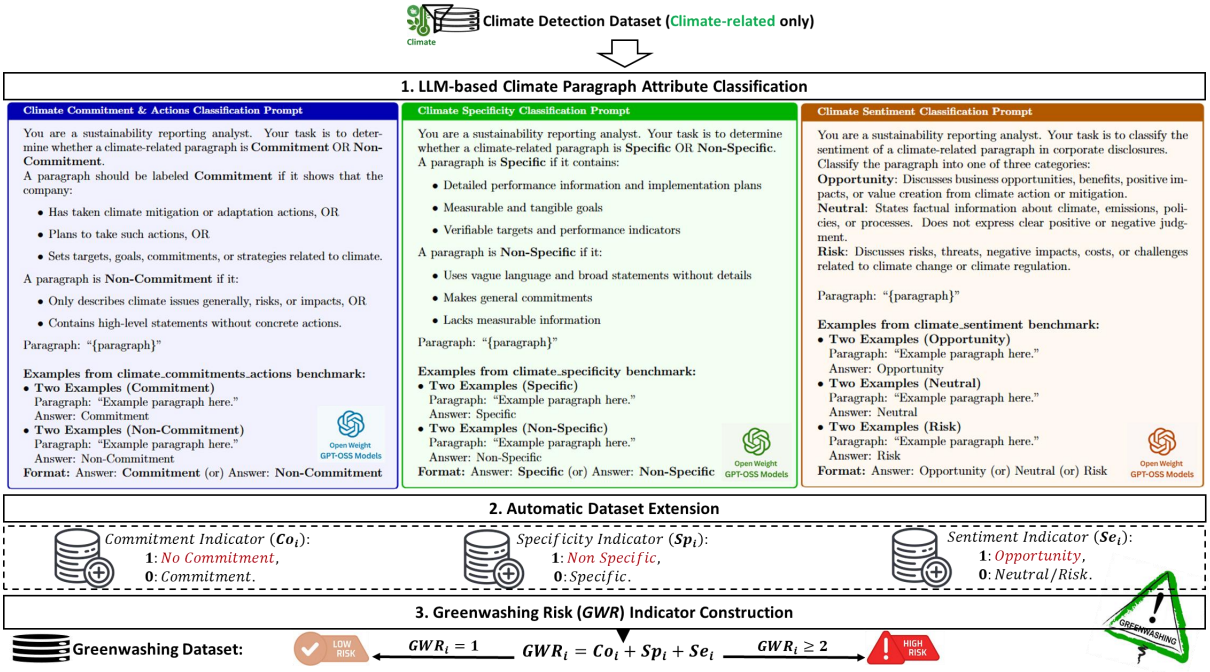

Figure 2: Overview of the greenwashing risk proxy construction pipeline based on prompt-based attribute classification, automatic dataset extension, and composite risk aggregation.

neutral if it is about facts without putting them into a positive or negative perspective. Lastly, it is labeled a risk if it mainly talks about business risks or the negative impacts of climate change."

Prompt responses are used to assign categorical labels for each dimension. This process yields automatic paragraph-level annotations that are consistent with existing benchmark formulations while enabling scalable extension to large, unlabeled corpora.

Automatic Dataset Extension. Using the prompting framework described above, we extend our automatically labeled climate dataset with commitment, specificity, and sentiment annotations. This procedure enables the creation of a large-scale paragraph-level dataset without requiring manual labeling.

Greenwashing Risk Proxy Construction. Building on prior work (Vinella et al., 2024), we operationalize greenwashing risk as a composite indicator reflecting three characteristic patterns commonly associated with misleading sustainability disclosures: absence of substantive commitments, vague language, and overly optimistic framing.

For each paragraph p_i, we define three binary indicators:

- **Commitment Indicator (Co_i):** 1 if the paragraph is classified as non-commitment, 0 otherwise.
- **Specificity Indicator (Sp_i):** 1 if the paragraph is classified as non-specific, 0 otherwise.
- **Sentiment Indicator (Se_i):** 1 if the paragraph exhibits optimistic framing (opportunity), 0 otherwise.

The Greenwashing Risk (GWR) proxy is computed as the unweighted sum:

$$GWR_i = Co_i + Sp_i + Se_i.$$

The score ranges from 0 to 3 and reflects the number of greenwashing-related linguistic signals present in a paragraph. We discretize the score into ordinal risk categories:

$$\text{Class}_i = \begin{cases} \text{Low Risk}, & \text{if } GWR_i = 1, \\ \text{High Risk}, & \text{if } GWR_i \geq 2. \end{cases}$$

Paragraphs with $\text{GWR}_i = 0$ do not exhibit any of the targeted greenwashing patterns. Although these paragraphs are climate-related, they are excluded from the risk classification task because our indicator is a textual proxy, not a definitive measure of compliance or deception. Labeling these paragraphs as fully compliant would require additional domain-specific supervision, such as expert audits or regulatory verification, which is beyond the scope of the current framework. Focusing on paragraphs with nonzero risk scores ensures the model learns from instances where textual patterns indicate potential greenwashing, avoiding overinterpretation of compliant disclosures.

As a result of this procedure, we obtain a paragraph-level dataset consisting of climate-related disclosures annotated with ordinal greenwashing risk proxy labels (Low Risk and High Risk). This automatically constructed dataset serves as the basis for benchmark augmentation and is subsequently used for downstream evaluation of greenwashing risk classification in subsection 4.2.

4. Experimental Evaluation

Using the datasets constructed in Section 3, we evaluate the effectiveness of our annotation pipeline through indirect benchmark augmentation, i.e., by expanding manually annotated benchmark datasets with automatically generated labels and measuring downstream performance gains. This setup allows us to assess whether the proposed annotations provide useful supervisory signals beyond simply increasing the training data size.

Our evaluation focuses on climate-related and greenwashing classification tasks using ClimateBERT, a domain-adapted transformer model initialized from DistilRoBERTa and further pretrained on a large corpus of climate-related research abstracts, corporate disclosures, and news articles. ClimateBERT has demonstrated strong performance across multiple climate-related classification benchmarks (Webersinke et al., 2022). Specifically, we assess whether incorporating automatically annotated data improves downstream classification performance on established climate-related benchmarks. We consider the following benchmark settings from the ClimateBERT benchmark suite:

- **Climate Detection**: binary classification of paragraph-level `climate relevance` dataset (Bingler et al., 2023b).
- **Greenwashing Risk Classification**: paragraph-level indicators derived from the `climate commitments actions` (Bingler et al., 2023a), `climate specificity` (Bingler et al., 2023d), and `climate sentiment` (Bingler et al., 2023c) datasets, which we transform into a unified greenwashing risk proxy (GWR) following the aggregation scheme described in Section 3.2.

These benchmark datasets were manually annotated by domain experts and analyzed in prior work (Bingler et al., 2024). Our greenwashing risk formulation follows Vinella et al. (2024).

We investigate the following research questions:

1. **RQ1.** Does augmenting benchmark datasets with automatically constructed annotations improve downstream performance, thereby demonstrating the informational value of the proposed pipeline?
2. **RQ2.** Does joint multi-task learning provide additional gains over separate single-task models?

Experimental Setup. We applied our pipeline to 30 sustainability reports spanning 10 industrial sectors, totaling 3,663 pages. After preprocessing, we obtained 19,133 paragraphs, of which 8,533 were associated with GRI climate-focused index entries. For climate labeling, the joint similarity threshold was set to $\Theta_1 = 0.35$, while $\Theta_2 = 0.01$ was used to identify non-climate paragraphs via the GRI 2 (General) filtering strategy described in Section 3.1. Each benchmark was expanded by approximately 50% of its original size. We fine-tuned ClimateBERT [1] on the augmented datasets and evaluated performance using Macro-averaged F1 metrics.

4.1. Climate Detection

For binary climate relevance classification, we use the `climate relevance` (Bingler et al., 2023b) dataset. The dataset consists of 1,300 training instances (1000 climate-related and 300 non-climate-related samples) and 400 test instances (320 climate-related and 80 non-climate-related samples). All samples are written in English and were collected from corporate disclosures. We augment the training set by adding 650 samples from our dataset, resulting in a combined training set of 1,950 samples (1648 climate-related and 302 non-climate-related samples). As shown in Table 1, augmenting the training data improves macro-averaged F1 by 1.2 points, improving overall macro-averaged F1, indicating better balance across classes.

Training Data	Macro-averaged F1 (%)
Original (1300 samples)	93.70
Combined (1950 samples)	**94.90**

Table 1: Climate detection performance on the test set (400 samples).

4.2. Greenwashing Risk Classification

Following the methodology described in Section 3.2, we construct paragraph-level greenwashing risk proxy labels by combining the outputs of three ClimateBERT downstream benchmarks: `climate commitments actions` (Bingler et al., 2023a), `climate specificity` (Bin-

[1] Training was performed using an effective batch size of 32 and learning rate of 2×10^{-5}. The model was trained for 5 epochs using the AdamW optimizer and a linear learning rate scheduler.

gler et al., 2023d), and climate sentiment (Bingler et al., 2023c). We apply the same process used for our main dataset to these benchmarks to derive a unified greenwashing risk proxy. Thus, benchmark and auto-labeled samples share a consistent labeling schema grounded in established expert annotations. Paragraphs that are classified as climate-related but for which all three indicators are zero (commitment, specificity, and optimistic sentiment indicators) are not assigned a risk label and are excluded from this analysis, as these cases require additional domain-specific resources for reliable assessment. As a result, the original dataset consists of 890 training instances (516 high-risk and 374 low-risk samples) and 280 test instances (178 high-risk and 102 low-risk samples). We evaluate the impact of dataset augmentation by comparing ClimateBERT trained on the original benchmark data with a model trained on the combined training dataset (1,500 samples: 927 high-risk and 573 low-risk), which includes our automatically labeled instances. Table 2 reports the results. As shown, augmenting the training data improves macro-averaged F1 by 2.4 points, indicating that the automatically constructed annotations provide informative supervisory signals for greenwashing risk modeling.

Training Data	**Macro-averaged F1 (%)**
Original (890 samples)	68.10
Combined (1500 samples)	**70.50**

Table 2: Single-task greenwashing risk classification performance on the test set (280 samples).

4.3. Multitask ClimateBERT Fine-Tuning

We perform multitask fine-tuning of ClimateBERT to jointly model climate detection and greenwashing risk classification (capturing climate commitments, specificity, and sentiment), leveraging shared patterns to improve generalization. To accommodate the multitask framework, we introduced an extended label scheme for greenwashing risk (GWR) classification. Since our training data is automatically generated and may contain noise, we adopt a prediction-based consistency constraint: when the model predicts a paragraph as NC, it is encouraged to assign the Not Consider label in the GWR task. This promotes cross-task consistency while retaining flexibility and reducing spurious predictions.

The overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{climate} + \mathcal{L}_{GWR} + \lambda \mathcal{L}_{consistency},$$

where $\lambda = 0.5$ controls the trade-off between task accuracy and cross-task consistency. This consistency mechanism is designed to improve robustness and mitigate cross-task inconsistencies, particularly in the presence of noisy automatically generated labels. Table 3 summarizes the macro-averaged F1 scores across both tasks.

Training Data	**Climate F1 (%)**	**GWR F1 (%)**
Original (1190 samples)	93.89	76.16
Combined (1802 samples)	**93.95**	**78.05**

Table 3: Multitask fine-tuning results on the test set (360 samples), reporting macro-averaged F1 scores for greenwashing risk and climate relevance classification.

4.4. Results: Analysis and Discussion

From the results summarized above, we derive empirical evidence addressing our two research questions.

RQ1. Across all experimental settings, dataset augmentation with automatically labeled paragraphs leads to consistent performance gains, demonstrating the quality of these data. For climate detection, augmenting the training data increases the macro-averaged F1 score from 93.7% to 94.9% (Table 1). For greenwashing risk classification, augmentation improves macro-averaged F1 from 68.1% to 70.5% in the single-task setting (Table 2) and from 76.16% to 78.05% in the multitask setting (Table 3).

RQ2. Multitask fine-tuning further improves greenwashing risk detection by jointly modeling climate relevance and risk indicators. Compared to the single-task GWR model (Table 2), multitask training increases macro-averaged F1 for greenwashing risk from 68.1% to 76.16% on the original dataset and from 70.5% to 78.05% on the combined dataset (Tables 2 and 3), while climate detection remains stable at 94% F1, demonstrating that joint optimization does not compromise the primary classification task. The multitask setting additionally includes non-climate paragraphs, which are assigned the Not Consider label for the greenwashing risk task; this accounts for the difference in dataset size relative to the single-task setting.

To account for the influence of additional non-climate samples labeled Not Consider, performance was also evaluated excluding this class, yielding consistent gains from 68.1% → 68.9% (original) and 70.5% → 71.79% (combined), confirming improvements are not solely driven by label distribution. This restricted evaluation provides a fairer comparison with the single-task setting, since it focuses only on comparable climate-

related instances. Class-wise analysis shows multi-task learning primarily enhances high-risk detection (76.6% → 79.7% original; 81.6% → 81.87% combined). In contrast, low-risk performance exhibits mixed trends, slightly decreasing on the original dataset (59.5% → 58.1%) but improving on the combined dataset (59.4% → 61.7%), indicating that joint modeling better captures subtle linguistic patterns associated with potential greenwashing, particularly when supported by automatically annotated data. To better understand model behavior, we conducted a brief qualitative analysis of representative predictions made by the multi-task ClimateBERT model. In particular, we examined cases where the model predicted a paragraph as climate-related and high-risk for greenwashing (NON-COMMITMENT: Active (1), NON-SPECIFIC: Active (1), SENTIMENT: Risk (0)). One illustrative test-set example reads:

> "*Ecological factors and environmental regulations for access to raw material deposits also create a degree of uncertainty. In some regions of the world, for example, in West Africa south of the Sahara, raw materials for cement production are so scarce that cement or clinker needs to be imported by sea. Rising transportation costs and capacity constraints in the port facilities can lead to an increase in product costs. Overall, we rate this as a low risk*."

This example illustrates a borderline case in which the paragraph contains environmental and regulatory language that is climate-relevant, yet does not express a concrete climate commitment or measurable mitigation action. The multitask model assigns a high-risk label because the statement remains relatively generic and non-specific, which is consistent with our proxy definition of potential greenwashing risk. More broadly, such cases highlight the difficulty of distinguishing between informative contextual discussion and disclosures that may remain vague or weakly substantiated from a climate accountability perspective.

5. Conclusion

This paper proposes a scalable, paragraph-level framework for aligning corporate sustainability disclosures with GRI indicators and TCFD pillars, integrating regulatory semantics with modern NLP architectures. By focusing on paragraph-level units, the framework addresses a critical granularity gap in current reporting practices, where climate-related information is dispersed across lengthy, unstructured documents and is difficult to retrieve and interpret systematically. Our results indicate that weakly supervised dataset augmentation improves classification robustness compared to models trained exclusively on manually annotated data. The combination of embedding-based similarity scoring with LLM-based validation enables efficient expansion of high-confidence training instances while controlling annotation noise. Fine-tuning ClimateBERT in multi-task settings further enhances generalization, indicating that climate relevance and linguistic greenwashing risk provide complementary supervisory signals. Nevertheless, given the relatively limited size of the evaluation sets, the observed improvements should be interpreted with appropriate caution. While the gains are consistent across all experimental settings, future work should validate these findings on larger manually verified benchmark datasets and through more extensive statistical significance analysis.

The introduction of a paragraph-level greenwashing proxy grounded in commitment intensity, specificity, and sentiment contributes a structured linguistic dimension to disclosure analysis. Unlike document-level ESG scoring approaches (e.g.,(Wood et al., 2025)), our framework captures within-document heterogeneity, enabling fine-grained identification of potentially symbolic versus substantive climate communication. A key strength of this approach lies in its alignment with regulatory requirements. While prior studies apply LLMs to ESG or climate-related text classification, they typically treat sustainability as a thematic category. In contrast, our approach leverages the official GRI-TCFD mapping as structured supervision, thereby reflecting the architecture of climate-related financial supervision. This structured alignment enhances interpretability and comparability, and increases the potential applicability of the framework to financial risk assessment. Furthermore, by reducing reliance on exhaustive manual annotation, the proposed pipeline lowers the cost barrier for large-scale climate disclosure monitoring. This is particularly relevant for banks, regulators, and institutional investors who must process vast volumes of unstructured textual data under evolving reporting standards (Wood et al., 2025). Future research may extend this framework by incorporating additional sustainability standards and developing longitudinal models to capture the temporal dynamics. Additional directions include improving the robustness of greenwashing risk estimation through richer expert-validated supervision, broader cross-sector evaluation, and more systematic qualitative error analysis. Overall, this study contributes to the intersection of climate finance and computational linguistics by providing a replicable, standard-aligned, and economically scalable methodology for AI-assisted climate disclosure analysis.

6. Bibliographical References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card.

Seyed Alireza Mousavian Anaraki, Danilo Croce, and Roberto Basili. 2025. Large language models for sustainability reporting: A systematic review and research agenda. *Sustainable Futures*, 10:101494.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.

M. Bronzini, C. Nicolini, B. Lepri, A. Passerini, and J. Staiano. 2024. Glitter or gold? deriving structured insights from sustainability reports via large language models. *EPJ Data Sci.*, 13(1):41.

Eduardo C Garrido-Merchán, Cristina González-Barthe, and María Coronado-Vaca. 2026. Fine-tuning climatebert transformer with climatext for the disclosure analysis of climate-related issues in corporates' financial and non-financial reports. *Neural Computing and Applications*, 38(1):12.

Anna Glazkova and Olga Zakharova. 2025. From data to grassroots initiatives: Leveraging transformer-based models for detecting green practices in social media. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 1–9, Tallinn, Estonia. University of Tartu Library.

Charlott Jakob, Vera Schmitt, Salar Mohtaj, and Sebastian Möller. 2024. Classifying sustainability reports using companies self-assessments. In *Future of Information and Communication Conference*, pages 547–557. Springer.

Agnieszka Janik and Adam Ryszko. 2025. Greenwashing in sustainability reporting: A systematic literature review of strategic typologies and content-analysis-based measurement approaches. *Sustainability*, 18(1):17.

Hyewon Kang and Jinho Kim. 2022. Analyzing and visualizing text information in corporate sustainability reports using natural language processing methods. *Appl. Sci.*, 12(11):5614.

A Kazakov, S Denisova, I Barsola, E Kalugina, I Molchanova, I Egorov, A Kosterina, E Tereshchenko, L Shutikhina, I Doroshchenko, et al. 2023. Esgify: Automated classification of environmental, social, and corporate governance risks. In *Doklady Mathematics*, volume 108, pages S529–S540. Springer.

Yao Li and Michael Rockinger. 2024. Unfolding the transitions in sustainability reporting. *Sustainability*, 16(2):809.

W. Moodaley and A. Telukdarie. 2023a. A conceptual framework for subdomain specific pretraining of large language models for green claim detection. *Eur. J. Sustain. Dev.*, 12(4):319.

W. Moodaley and A. Telukdarie. 2023b. Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review. *Sustainability*, 15(2):1481.

Seyed Alireza Mousavian Anaraki, Danilo Croce, and Roberto Basili. 2025a. Automatic GRI-SDG annotation and LLM-based filtering for sustainability reports. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 775–784, Cagliari, Italy. CEUR Workshop Proceedings.

Seyed Alireza Mousavian Anaraki, Danilo Croce, and Roberto Basili. 2025b. Unsupervised sustainability report labeling based on the integration of the GRI and SDG standards. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 151–162, Vienna, Austria. Association for Computational Linguistics.

Hui Qian Ngee, Asha Ganesh, Muhammad Aizat Noor Azmi, Tiong Yew Tang, Muaadh Mukred, Fathey Mohammed, and Adi Affandi Bin Ahmad. 2024. Environmental. social and governance (esg) scores automation in global reporting initiative (gri) with natural language processing. In *2024 7th International Conference on Internet Applications, Protocols, and Services (NETAPPS)*, pages 1–7. IEEE.

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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Adv. Neural Inf. Process. Syst.*, 33:16857–16867.

Avalon Vinella, Margaret Capetz, Rebecca Pattichis, Christina Chance, Reshmi Ghosh, and Kai-Wei Chang. 2024. Leveraging language models to detect greenwashing.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text.

Katherine Wood, Chaehyun Pyun, and Hieu Pham. 2025. Beyond green labels: assessing mutual funds' esg commitments through large language models. *Finance Research Letters*, 74:106713.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

Faturahman Yudanto, Yunita Sari, and Maeve Zahwa Adriana Crown Zaki. 2024. Climate-NLI: A model for natural language inference and zero-shot classification on climate-related text. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 600–608, Tokyo, Japan. Tokyo University of Foreign Studies.

Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, ZongXiong Lei, Zihan Zeng, Shiming Yang, Hongxiang Tong, Lei Xiao, and Wenwen Zhou. 2025. Esgreveal: An llm-based approach for extracting structured data from esg reports. *Journal of Cleaner Production*, 489:144572.

7. Language Resource References

Bingler et al. 2023a. *ClimateBERT Climate Commitments Actions Dataset*. ClimateBERT Project. HuggingFace Dataset Repository. Dataset of climate-related commitments and actions annotated from corporate disclosures. Licensed under CC BY-NC-SA 4.0.

Bingler et al. 2023b. *ClimateBERT Climate Detection Dataset*. ClimateBERT Project. HuggingFace Dataset Repository. Expert-annotated dataset for detecting climate-related paragraphs in corporate disclosures. Licensed under CC BY-NC-SA 4.0.

Bingler et al. 2023c. *ClimateBERT Climate Sentiment Dataset*. ClimateBERT Project. HuggingFace Dataset Repository. Paragraph-level climate sentiment classification dataset. Licensed under CC BY-NC-SA 4.0.

Bingler et al. 2023d. *ClimateBERT Climate Specificity Dataset*. ClimateBERT Project. HuggingFace Dataset Repository. Paragraph-level climate specificity classification dataset. Licensed under CC BY-NC-SA 4.0.