Greench-v1: distilling SLMs on Greenwashing Detection

Simona Scala, Federico Raspanti, Irem Demirtas,
Alessandro Pietro Bardelli, Marilena Di Bari
Michele Filannino
Prometeia spa
Piazza Trento e Trieste, 3, 40137 Bologna
{simona.scala, federico.raspanti, irem.demirtas,
alessandro.bardelli, marilena.dibari, michele.filannino}@prometeia.com

Abstract

Validating greenwashing claims in environmental, social, and governance (ESG) reports relies heavily on costly and inconsistent manual review. To address this, this paper introduces Greench-v1, a low-latency small language model (based on Qwen3-4B) that screens ESG text at the paragraph level. The model outputs a three-way classification (Greenwashing Alert, No Greenwashing, Not Relevant) paired with a concise, paragraph-grounded rationale to assist human auditors in triage and validation. The system was trained on a custom dataset of roughly 2,000 paragraphs, adapted from the ClimateBERT corpus. This dataset mitigates class imbalance through controlled paraphrasing of rare positive instances and uses GPT-4o to generate evidence-based justifications. Four training regimes were evaluated: (i) Hard distillation: Supervised fine-tuning on teacher-generated outputs. (ii) Soft distillation: Training the student to match the temperature-scaled logits of a domain-specialized Qwen3-14B teacher. (iii) Group Relative Policy Optimization (GRPO): Reward-based updates driven by exact-match alert generation. (iv) Hybrid GRPO: GRPO initialized from the hard-distilled checkpoint. Distillation and efficient policy optimization significantly improved performance over untuned baselines. Soft distillation and GRPO achieved the strongest results, increasing the "Greenwashing Alert" weighted F1-score by 36.7% and 49.0%, respectively, resulting in a deployable tool for screening large volumes of ESG narratives.

Keywords: ESG reporting, greenwashing detection, knowledge distillation, policy optimization, small language models.

1. Introduction

Over the past decade, the progressive evolution of the European sustainability regulatory framework—together with the growing demand for sustainable finance—has prompted banks, as well as other financial and non-financial institutions, to systematically integrate **Environmental, Social and Governance (ESG) factors** into their strategies, product offerings, risk management frameworks, and operating models.

The increasing relevance of sustainability thus materializes across multiple dimensions, reflecting regulatory developments, heightened stakeholder expectations, and the evolving role of financial institutions in supporting a just and orderly transition to a more sustainable economy. Among these dimensions, the safeguarding of the reputational sphere has assumed particular importance. This entails the continuous monitoring of alignment between publicly disclosed commitments (e.g. sustainability reports, climate transition plans, green bond frameworks, annual integrated reports, product-level ESG factsheets, etc.) and actual business practices, with the objective of mitigating exposure to greenwashing risk. **Greenwashing** refers to the practice whereby an institution misrepresents, exaggerates, or selectively discloses information regarding the environmental or sustainability-related characteristics of its products, services, or overall strategy, thereby creating a misleading perception of ESG alignment. Such practices may take the form of vague or unsubstantiated claims, incomplete or biased transparency, or the overemphasis of isolated sustainable initiatives while disregarding broader adverse environmental or social impacts. Beyond regulatory and legal implications, greenwashing poses significant reputational risks and undermines stakeholder trust and the credibility of sustainable finance frameworks.

In response to these risks, increasing attention has been devoted to the development of methodologies for greenwashing detection, particularly through the analysis of ESG-related disclosures, such as sustainability reports and other non-financial statements.

From a technological perspective, recent advancements in Natural Language Processing (NLP) have significantly enhanced detection capabilities. In particular, **Generative AI techniques, such as Large Language Models (LLMs) and Small Language Models (SLMs)** have progressively replaced traditional vocabulary-based or purely statistical approaches, enabling more nuanced, context-aware, and semantically rich analyses of textual content. These advancements have strengthened

Proceedings of the 2nd Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology) @ LREC 2026, pages 79–86

institutions' ability to identify inconsistencies, overly generic claims, and potential misalignments between narrative disclosures and measurable sustainability performance.

2. Related Works

2.1. NLP-Based Methodologies for Evaluating ESG Disclosures

The operationalization of greenwashing detection has evolved significantly from binary assessments of truthfulness to multidimensional evaluations of **narrative strategy and communicative decoupling**.

Early NLP approaches relied heavily on simple lexicons to analyze tone and readability, finding that environmental violators often utilize more positive, verbose, yet less readable language to deflect stakeholder attention (Gorovaia and Makrominas, 2025). The field subsequently advanced with transformer-based architectures, notably **ClimateBERT**, which outperforms general-purpose models by pre-training on domain-specific climate corpora (Webersinke et al., 2021).

Building on these foundation models, scholars generally agree that detecting greenwashing requires analyzing the structural and stylistic cues of a text rather than relying solely on keyword frequencies. For instance, Bingler et al. (2024) developed the "Cheap Talk Index" using ClimateBERT to identify non-specific climate commitments. To capture this complexity, researchers have proposed composite metrics like the *Greenwashing Severity Index (GSI)* and the *Green Authenticity Index (GAI)*, which combine sentiment analysis, TF-IDF weighting, and topic modeling to quantify the divergence between a firm's reported narratives and independent external evidence Sudro and Mukhopadhyay (2025). Despite this consensus on the utility of NLP, methodologies diverge on validation: some researchers rely purely on internal linguistic characteristics like hedging and vagueness (Livytska, 2019), whereas others argue that textual signals must be benchmarked against external performance data, such as RepRisk incidents, to conclusively prove substantive decoupling.

2.2. The advent of Large Language Models

A fundamental gap identified across the current literature is the **absence of comprehensive, gold-standard datasets** containing verified cases of greenwashing, largely due to the subjective, ambiguous, and legally sensitive nature of the phenomenon (Calamai et al., 2025). To circumvent this bottleneck, recent studies have increasingly employed Large Language Models (LLMs) to synthesize training data or act as zero-shot evaluators. Birti et al. (2025) successfully demonstrated that augmenting manual annotations with LLM-generated synthetic data, such as controlled paraphrasing of ESG statements, significantly improves the classification accuracy of downstream models.

However, relying on massive, general-purpose LLMs introduces profound vulnerabilities into the auditing pipeline. Chuang et al. (2025) empirically demonstrated that LLMs can be weaponized by corporations to seamlessly generate highly convincing greenwashed responses that evade standard detection mechanisms, unless strict accuracy constraints are externally imposed. These vulnerabilities highlight a critical gap in the literature: massive, **black-box LLMs are too resource-intensive, unpredictable, and easily manipulated** to serve as reliable, large-scale financial auditing tools.

2.3. The Shift Toward more efficient and explainable Models

To achieve the deep reasoning capabilities of massive LLMs without their computational overhead and unpredictability, the literature is pivoting toward **Small Language Models (SLMs)** enhanced via **Knowledge Distillation (KD)** and reinforcement learning to democratize language evaluation. KD addresses the limitations of standard training by transferring the continuous probability distributions (logits) of a massive teacher model into a compact student model, allowing the SLM to internalize complex domain nuances while remaining accessible to researchers with limited compute resources. Recent advancements emphasize **"Distilling Step-by-Step,"** wherein student models are trained on natural language chain-of-thought rationales generated by the teacher, enabling them to outperform much larger models while utilizing fewer parameters (Hsieh et al., 2023). Furthermore, post-training alignment techniques, such as **Group Relative Policy Optimization (GRPO)**, allow these models to learn from verifiable, multi-objective rewards, ensuring that their evaluation outputs remain factually grounded, structurally sound, and scalable for reproducible research without the prohibitive costs of traditional reinforcement learning.

3. Greench-v1

The aforementioned research issues (the necessity for multi-dimensional linguistic evaluation, the critical bottlenecks of data scarcity and LLM vulnerabilities, and the promise of KD-optimized SLMs) directly frame the specific problem addressed by **Greench-v1**. While current approaches either rely on simplistic, static proxies or brittle, computation-

ally expensive LLMs that lack the transparency required for rigorous financial assurance, *Greench-v1* addresses these limitations through a compact, locally deployable architecture that enables low-latency inference, full training transparency, and structured label-and-rationale outputs suited for human-in-the-loop ESG auditing pipelines.

Processing Pipeline

The deployed model utilized a custom soft-distillation trainer. Additionally, the prompt structure was optimized by explicitly defining greenwashing in the system prompt. The compiled Greench-v1 model is operationalized via a web-based interface that strictly orchestrates document processing from ingestion to output generation. The procedural workflow is as follows:

1. Input Ingestion: A user uploads a target sustainability document (in PDF format) into the system interface.

2. Execution and Chunking: Upon user initiation, the system performs Optical Character Recognition (OCR) and segments the entire document into discrete paragraphs.

3. Iterative Processing: The core algorithm analyzes the text sequentially, outputting the results paragraph by paragraph to the interface.

4. Classification Formulation: For every processed paragraph, the model computes a classification and corresponding justification. It assigns "Not relevant" to text lacking environmental claims, "No greenwashing" to substantiated claims (e.g., factual references to operational certifications), and triggers a "Greenwashing alert" when identifying broad, unmeasured environmental commitments lacking concrete action (Fig. 1)

5. Synthesis and Export: Once paragraph-level iteration concludes, the system aggregates a summary of the analysis at the bottom of the interface. Finally, the granular results, including text chunks, labels, and justifications, can be exported into standard, machine-readable file formats (such as CSV or JSON) for external auditing or reproducibility (Fig 2).

4. Data Collection and Preprocessing

4.1. Primary Dataset Construction

The dataset construction commenced with the `climatebert/climate_specificity` dataset, a binary classification task determining whether a given climate-related paragraph is specific or non-specific in nature (Webersinke et al., 2021), yielding 1,320 paragraphs. These instances were mapped to formulate an initial dataset categorized by a binary classification schema into “POSSIBLE GREENWASHING” and “NO GREENWASHING” labels. To incorporate a null class, 660 paragraphs (one third of the total) explicitly classified as non-environmental claims were sampled from the `climatebert/environmental_claims` dataset, which provides a binary classification task identifying whether a given sentence constitutes an environmental claim or not, and designated as “NOT RELEVANT” (Webersinke et al., 2021).

To mitigate class imbalance, the underrepresented “POSSIBLE GREENWASHING” class was augmented via controlled paraphrasing. Specifically, each positive instance was paraphrased three times using GPT 4.1, to expand the minority class distribution and match the distribution of the null and negative classes. Subsequently, evidence-based textual justifications for each sample were synthesized utilizing GPT-4.1, which generated a rationale strictly conditioned on the source paragraph and its corresponding label. The final preprocessing pipeline yielded a curated corpus of approximately 2,000 paragraphs, wherein each entry comprises the source text, a discrete three-way classification label, and an accompanying natural language justification.

The resulting corpus was partitioned into training and evaluation sets via a stratified 80/20 split, yielding approximately 1,600 paragraphs for training and 400 for evaluation, with class proportions preserved across both subsets.

4.2. Supplementary ESG Corpus for Hard Distillation

In addition to this primary corpus, a supplementary dataset was constructed exclusively for the Hard Distillation + GRPO training paradigm described in Section 5.1. This dataset was assembled from 30 real-world ESG reports, segmented at the paragraph level. Each paragraph was annotated by GPT-4.1 according to the same three-way classification schema (“POSSIBLE GREENWASHING”, “NO GREENWASHING”, “NOT RELEVANT”), accompanied by a corresponding natural language justification. The objective of this corpus is to expose the model to a broad distribution of authentic ESG discourse prior to policy optimization, thereby providing domain familiarization at scale before the reward-driven fine-tuning stage.

Field	Content
Source Text	*In 2022, Mango joined the network of the Ellen MacArthur Foundation, an international charity dedicated to promoting the circular economy. This vision aims to eliminate waste and pollution, encourage the circulation of products and materials, and regenerate nature, with the goal of transforming the economic system towards a more sustainable one.*
Label	POSSIBLE GREENWASHING
Justification	This paragraph presents a general statement about Mango's association with the Ellen MacArthur Foundation. However, there is no specific mention of concrete actions taken nor detailed information regarding how these objectives have been implemented or achieved. The text relies on broad, unsubstantiated references to organisational commitments rather than tangible outcomes.

Figure 1: Example paragraph-level output from Greench-v1 on a Mango ESG report, flagged as POSSIBLE GREENWASHING due to the lack of concrete, measurable environmental commitments.

Category	Count
Total paragraphs	29
Greenwashing alerts	6
Not relevant	11
Clean paragraphs	12

Figure 2: Example summary statistics produced by Greench-v1 after processing a Mango ESG report. Results can be exported in CSV or JSON format for external auditing or reproducibility.

5. Methodology

5.1. Experiments

All training and inference experiments were conducted on a single instance (24 vCPUs, 220 GB RAM, 64 GB disk, equipped with one NVIDIA A100 40 GB GPU). Inference on the held-out evaluation set of approximately 400 paragraphs completed in under 5 minutes.

To rigorously select the optimal learning strategy for the Greench-v1 architecture, four distinct training paradigms were investigated (Fig.3).

First, a **hard distillation** approach was implemented, wherein the Qwen3-4B model was supervised directly on the ground-truth targets, encompassing both the discrete classification labels and their associated textual justifications.

Second, a **soft distillation** methodology was examined. This paradigm involved the initial fine-tuning of a Qwen3-14B teacher model, followed by the optimization of the Qwen3-4B student using a composite loss function. This function integrates the cross-entropy loss over the gold tokens with the **forward Kullback-Leibler divergence** computed between the temperature-scaled logits of the teacher and the student. The formulation incorporates hyper-parameters α and T to modulate the relative weighting of the teacher's distribution against the ground-truth answers, and to adjust the distributional entropy of the teacher's signals. However, implementing this approach incurred a soft-distillation adaptation cost, as the custom script required time to be adapted from vision models to transformers.

Third, **Group Relative Policy Optimization (GRPO)** was evaluated using a discrete reward mechanism. The reward function strictly assigned $+1$ for an exact syntactic match with the ground-truth label and $+0$ otherwise. Despite this straightforward assignment, managing GRPO complexity was challenging, as defining an appropriate and stable reward function proved difficult. Throughout the GRPO training phase, the model decoded 10 distinct candidate sequences per optimization step, with the policy gradient algorithm updating the network parameters to maximize the averaged expected reward across the generated group.

Finally, a hybrid **Hard Distillation + GRPO** paradigm was investigated to assess whether domain familiarization prior to policy optimization could further improve classification performance. In this regime, Qwen3-4B was first subjected to hard distillation on the supplementary ESG report corpus described in Section 4.2, exposing the model to a large and diverse distribution of authentic ESG narratives. The resulting checkpoint was subsequently used to initialize GRPO training on the primary dataset, applying the same exact-match reward function as in the standalone GRPO condition. The rationale underlying this two-stage procedure is that broad domain exposure during the distillation phase may yield a more favorable **parameter initialization** for reward-driven fine-tuning, potentially accelerating convergence and improving robustness.

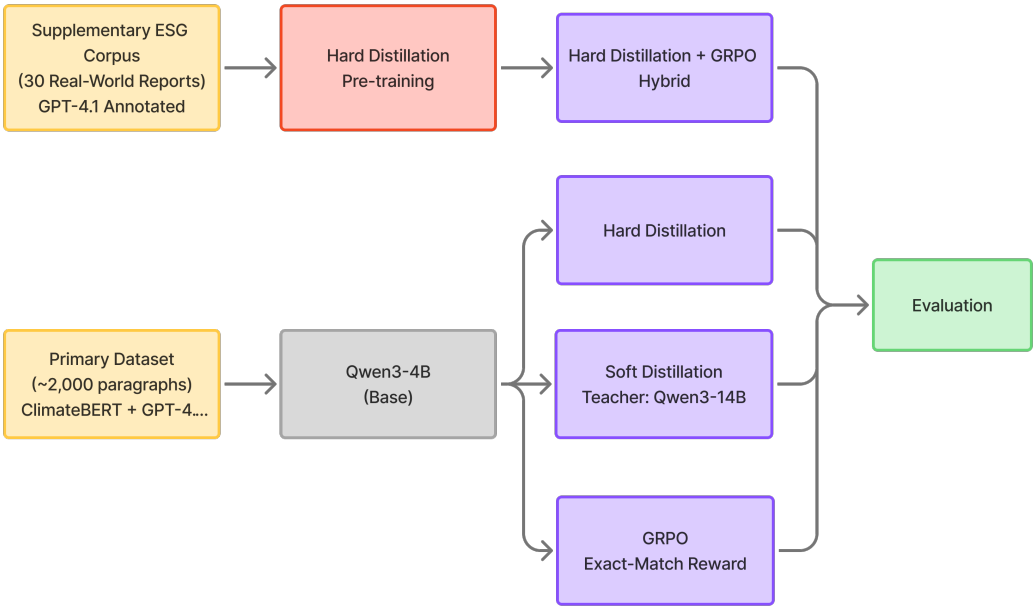

Figure 3: Overview of the four training paradigms evaluated for Greench-v1: Hard Distillation, Soft Distillation (with Qwen3-14B as teacher), GRPO with exact-match reward, and a Hybrid Hard Distillation + GRPO approach initialized from the supplementary ESG corpus checkpoint.

6. Results

6.1. Evaluation Metrics

To assess model performance across both output dimensions, two complementary metrics were employed. For the discrete classification task, the **weighted F1-score** was adopted to evaluate Greenwashing Alert predictions.

For the free-text justification output, **BERTScore** (Zhang et al., 2020) was used to measure the semantic similarity between the model-generated rationale and the reference justification. BERTScore leverages contextual embeddings from a pre-trained language model to compute token-level similarity via greedy matching, producing precision, recall, and F1 estimates that capture meaning beyond **surface-level lexical overlap**. The F1 variant of BERTScore is reported throughout.

For the discrete classification task, the **weighted F1-score** was adopted to evaluate Greenwashing Alert predictions. Although the evaluation set is approximately balanced across classes, weighted F1 was retained for consistency with early experimental runs conducted on the original imbalanced corpus; under balanced conditions, the weighted and macro-averaged F1-scores are functionally equivalent.

6.2. Quantitative Evaluation

Table 1 reports the Greenwashing Alert weighted F1-score and the Justification F1-score for all evaluated baselines and distilled model variants.

The results reveal a consistent and substantial **performance gap** between the untrained baselines and the distilled model variants across both evaluation dimensions. Among the baselines, the untuned Qwen3-4B achieves the highest Greenwashing Alert weighted F1-score of 0.49, which is notably counter-intuitive given that Qwen3-32B—a considerably larger model—underperforms at 0.45. This suggests that **raw parameter count does not directly confer advantage** in this domain-specific task, and that a smaller model already fine-tuned for instruction-following may be more amenable to zero-shot greenwashing classification. This suggests that **raw parameter count does not directly confer advantage** in this domain-specific task; the marginal gap of 0.04 between the two models is small enough to be attributable to stochastic variation in decoding rather than a systematic capability difference, and both models should be regarded as equivalent zero-shot baselines. `claude-4.5-sonnet` records the weakest classification score (0.34) despite matching Qwen3-4B on the Justification F1 metric (0.62), indicating a dissociation between its capacity to generate coherent ratio-

Model	Greenwashing Alert (Weighted F1)	Justification (F1)
Baselines		
Qwen3-4B	0.49	0.62
Qwen3-32B	0.45	0.58
claude-4.5-sonnet	0.34	0.62
Distilled Models		
Qwen3-4B-hard-distillation	0.54	0.59
Qwen3-4B-soft-distillation	0.67	**0.64**
Qwen3-4B-grpo	**0.73**	0.62
Qwen3-4B-hard-distillation + grpo	0.60	0.62

Table 1: Performance of baseline and distilled models on the Greenwashing Alert weighted F1-score and Justification F1-score. Bold entries indicate the best result per column.

nales and its ability to reliably assign the correct alert label under zero-shot conditions. Indeed, we find that `claude-4.5-sonnet` was too eager at classifying paragraphs as potential greenwashing.

Across the distilled variants, all four training paradigms improve upon the best baseline on the Greenwashing Alert metric, confirming that **task-specific supervision** is a necessary condition for reliable greenwashing triage. Hard distillation yields a modest improvement to 0.54, demonstrating that supervised fine-tuning on teacher-generated gold labels alone is insufficient to fully capture the distributional nuances of the classification task. The most pronounced gains are attributable to soft distillation and GRPO. Soft distillation achieves a weighted F1-score of 0.67—a relative improvement of 36.7% over the best baseline—and simultaneously records the highest Justification F1-score of 0.64, suggesting that exposure to the teacher's full **token-level probability distribution** yields richer, more transferable representations that benefit both classification accuracy and rationale quality. GRPO attains the strongest Greenwashing Alert score of 0.73, corresponding to a 49.0% relative gain over the best baseline, underscoring the effectiveness of reward-driven policy optimization in sharpening the model's sensitivity to the exact syntactic structure of alert labels. However, GRPO does not improve the Justification F1-score beyond 0.62, implying that a **binary exact-match reward** is sufficient to steer classification behavior but does not incentivize qualitative improvements in the accompanying rationale.

The Hard Distillation + GRPO hybrid achieves a Greenwashing Alert weighted F1-score of 0.60 and a Justification F1-score of 0.62. While this represents a meaningful improvement over the hard distillation baseline (0.54), it falls short of the standalone GRPO result (0.73). This outcome suggests that initializing from a checkpoint trained on the large ESG corpus does not provide a more favorable starting point for policy optimization than the default instruction-tuned initialization. A plausible explanation is that hard distillation on GPT-4.1-annotated ESG reports, while broadening domain coverage, may simultaneously introduce **labeling noise** or stylistic biases that partially interfere with the reward signal during the subsequent GRPO phase. Nonetheless, the hybrid approach does match the standalone GRPO on justification quality (0.62), indicating that domain pre-exposure does not degrade rationale generation.

6.3. Qualitative Evaluation

To complement the quantitative evaluation, we examine representative predictions to characterize the system's behaviour across correct and incorrect classifications.

True Negative. The model correctly assigns No Greenwashing Detected to a paragraph discussing climate-related physical risks, such as the expected southward extension of cyclones in Australia and their implications for asset pricing. The justification correctly identifies that the text offers factual, risk-oriented disclosure without any attempt to portray the institution as environmentally responsible, demonstrating that the model has learned to distinguish risk reporting from sustainability marketing.

True Positive. The model correctly flags a paragraph in which an investor states a general intention to contribute to a circular economy, citing their network, scale, and influence, without providing any measurable targets, concrete actions, or verified outcomes. The generated justification accurately identifies the reliance on aspirational language and the absence of substantiated commitments as the key markers of greenwashing.

False Positive (Type I Error). The model incorrectly raises a Greenwashing Alert for a paragraph describing a $25 million investment in electric vehicle charging infrastructure, including specific pilots in Minnesota and plans for fleet conversion.

While the target label is NO GREENWASHING DETECTED, the model penalizes the use of forward-looking language ("expect to expand") and the absence of verified outcome metrics. This suggests a systematic tendency to over-flag paragraphs that mix concrete actions with future-oriented commitments.

False Positive (Type I Error). A second false positive arises on a paragraph reporting dedicated ESG headcount across Finance, Communications, and Risk departments, including ten staff working on climate-related risk methodologies. Despite the specificity of the staffing figures, the model flags the paragraph due to the lack of measurable outcomes or verified results linked to these activities. This indicates that the model may insufficiently reward operational transparency in the absence of explicit performance metrics.

7. Conclusions

This paper presented Greench-v1, a lightweight, proprietary small language model for paragraph-level greenwashing detection in ESG disclosures. Empirical results demonstrate that task-specific supervision via knowledge distillation and policy optimization substantially outperforms zero-shot baselines, with soft distillation and GRPO yielding relative improvements of 36.7% and 49.0% on the Greenwashing Alert weighted F1-score, respectively.

Several properties of Greench-v1 make it particularly well-suited for deployment in real-world ESG auditing pipelines. First, its compact 4B-parameter architecture enables low-latency inference at scale, rendering it feasible for organizations lacking access to high-performance compute infrastructure. Second, as a proprietary, locally deployable model, Greench-v1 is not subject to undisclosed behavioral updates or capability drift associated with commercial API-based systems, ensuring reproducibility and auditability over time. Third, full transparency over the training corpus, derived from the ClimateBERT dataset with controlled augmentation, permits principled assessment of the model's domain coverage and potential biases, a prerequisite for regulatory-grade financial assurance. Finally, the modular paragraph-level interface, with its structured label-and-rationale output schema, serves a dual purpose. As a pre-publication writing assistant, it enables ESG report editors to iteratively refine paragraph wording, reduce exposure to reputational and regulatory risk, and ensure claims are substantiated prior to disclosure. As a post-publication triage component, it integrates within broader, multi-stage ESG analysis systems, including document-level aggregation pipelines or human-in-the-loop auditing workflows.

From a broader financial perspective, Greench-v1 represents a methodologically novel contribution to the intersection of NLP and financial compliance. Unlike general-purpose LLMs deployed as black-box evaluators, Greench-v1 introduces a transparent, auditable, and resource-efficient framework for automated narrative scrutiny that directly addresses the operational and regulatory constraints faced by financial institutions. The combination of knowledge distillation and policy optimization offers a replicable methodology for developing domain-specialized compliance tools beyond greenwashing detection. It is applicable, for instance, to the screening of prospectuses, loan documentation, or product-level ESG factsheets for regulatory adherence. In an environment of increasingly stringent disclosure requirements, the ability to deploy locally auditable, low-latency models at scale constitutes a meaningful advance over purely manual or API-dependent review processes. Greench-v1 thus demonstrates that compact, distilled language models can serve as credible first-line screening tools within institutional risk management and compliance workflows, reducing both the cost and latency of ESG-related due diligence.

Qualitative analysis reveals that the primary failure mode of the best-performing model consists of false positives arising from paragraphs that combine concrete actions with forward-looking language or report operational transparency without explicit performance metrics. Addressing this limitation is a key priority for future development.

Future work should investigate the extension of the reward function in GRPO to encompass rationale quality metrics, as well as the evaluation of Greench-v1 on external, independently curated greenwashing benchmarks to assess generalization beyond the ClimateBERT domain. Additionally, extending the framework to multimodal greenwashing detection represents a promising avenue, as ESG documents frequently embed visual elements—such as images, graphs, and charts—that may convey or obscure sustainability claims independently of the accompanying text; incorporating such modalities could yield a more comprehensive assessment of disclosure integrity.

8. Acknowledgements

The authors would like to thank our colleagues Lorenzo Prosperi and Michele Cimino for their valuable insights and constructive feedback on the overall framework.

9. Bibliographical References

ClimateBERT: A pretrained language model for climate-related text.

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 Birti, F Osborne, and A Maurino. 2025. Optimizing large language models for esg activity detection in financial texts. arxiv. *arXiv preprint arXiv:2502.21112*.

Tom Calamai, Oana Balalau, Théo Le Guenedal, and Fabian M Suchanek. 2025. Corporate greenwashing detection in text–a survey. *arXiv e-prints*, pages arXiv–2502.

Marianne Chuang, Gabriel Chuang, Cheryl Chuang, and John Chuang. 2025. Judging it, washing it: Scoring and greenwashing corporate climate disclosures using large language models. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 17–31, Vienna, Austria. Association for Computational Linguistics.

Nina Gorovaia and Michalis Makrominas. 2025. Identifying greenwashing in corporate-social responsibility reports using natural-language processing. *European Financial Management*, 31(1):427–462.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Inna Livytska. 2019. The use of hedging in research articles on applied linguistics. *Journal of language and cultural education*, 7(1):35–53.

Protima Nomo Sudro and Shreya Mukhopadhyay. 2025. Greenwashing detection with causal explanation: A novel multi-layered approach. In *Women in Machine Learning Workshop@ NeurIPS 2025*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

10. Language Resource References

Webersinke, Nicolas and Kraus, Mathias and Bingler, Julia Anna and Leippold, Markus. 2021.