Leveraging AI and ML Frameworks to Detect and Mitigate Greenwashing in Corporate ESG Reporting

1 [Ashwika Shivalkar], 2 [Diya R Nair], 3 [Pankhuri Govila]

Abstract

Corporate sustainability disclosures increasingly rely on algorithmic and language-based systems to communicate environmental performance, raising concern that automated reporting practices may obscure a widening gap between what firms claim and what they actually do, a practice commonly termed greenwashing. This study examines whether artificial-intelligence and machine-learning methods can reliably detect this gap, and what form an effective detection approach should take. Adopting a design-oriented research strategy, the study first synthesis literature spanning language-based, quantitative, visual, and legal-governance approaches to detection, then empirically tests several of the field's central claims using a secondary panel of firms drawn from three carbon-intensive sectors over a fifteen-year period. Descriptive statistics and cluster-adjusted significance testing were combined with supervised classification models evaluated on firms entirely unseen during training. The analysis confirms a robust and consistent discrepancy between reported sustainability performance and underlying emissions trends, but shows that the size of this discrepancy depends heavily on which performance benchmark is used, and that predicting greenwashing from disclosure content and self-reported scores alone performs only marginally better than chance once definitional and evaluation artefacts are corrected. The findings suggest that textual and self-reported evidence, taken alone, cannot support reliable automated detection, and that credible systems must instead combine narrative analysis with independently verified operational evidence. The study concludes by outlining the requirements for such a combined architecture, to be developed and tested in subsequent research.

Keywords: corporate sustainability disclosure; automated greenwashing detection; disclosure–performance gap; carbon-intensive industries; panel data analysis; predictive validity; algorithmic accountability

Table of Contents

Abstract 1
1. Introduction 4
1.1Background and Context 4
1.2 Importance of the Topic 4
1.3 Review of Existing Literature 5
1.4 Research Gap 6
1.5 Research Problem 6
1.6 Research Questions and Objectives 6
1.7 Significance of the Study 7
2. LITERATURE REVIEW 8
2.1 Executive Summary 8
2.2 NLP and Transformer-Based Frameworks 8
2.3 Quantitative Modeling and the Disclosure-Performance Gap (DPG) 9
2.4 Multimodal Auditing and the Greenwashing Paradox 9
2.5 Legal Liability, Algorithmic Accountability, and Governance 9
2.6 Research Gaps and the Proposed Contribution 10
2.7 Conclusion 11
3. METHODOLOGY 11
3.1 Research Design 11
3.2 Nature of Research 12
3.3 Sample & its Technique 13
3.3.1 Literature Sample 13
3.3.2 Dataset Sample 13
3.3.3 Data Collection Tool 13
3.3.4 Variables Used 14
3.4 Ethical Considerations 15
3.5 Limitations 16
4. DATA ANALYSIS AND INTERPRETATION 17
4.1 Executive Summary 17
4.2 Dataset and Objective 17
4.3 Greenwashing Prevalence and the Talk–Walk Discrepancy 17
4.4 Crosswashing, Net-Zero Paradox, and Predictive Modelling 19
4.5 Cross-Jurisdictional Findings and Data Limitations 21
4.6 Implications for Phase 3 and Conclusion 21
5. Discussion 22
5.1 The Central Talk–Walk Discrepancy 22
5.2 The Rating-Lens Dependency Problem 23

5.3 The Crosswashing Hypothesis: An Unexpected, Reversed Finding 23
5.4 The Net-Zero Commitments Paradox 24
5.5 Cross-Jurisdictional Divergence: Detection Sensitivity over Prevalence 24
5.6 Predictive Performance and the Limits of the Modelling Exercise 25
5.7 Implications for the Proposed Phase 3 Architecture 25
5.8 Caution on Causal Language and Generalisability 25

6. Conclusion 26

6.1 Restatement of the Research Focus 26
6.2 Major Findings 26
6.3 Overall Interpretation 26
6.4 Contribution 26
6.5 Implications 27
6.6 Limitations 27
6.7 Future Research 27
6.8 Concluding Remarks 28
6.9 Policy Recommendations 28
References 30

LIST OF TABLES

Table 1: Empirical Limitation in Current State-of-the-Art and Proposed Architectural/Methodological Contribution 10
Table 2: Operational Definitions of Independent and Derived Variables 14
Table 3: Country-Level Greenwashing Prevalence and Sample Distribution 18
Table 4: Original Predictive Model Performance (5-Fold Cross-Validation with Definitional Leakage) 19
Table 5: Comparison of Original Headline Results and Corrected Unseen-Firm Test Performance 20
Table 6: Firm Size Impact: Empirical Findings and Comparison to Prior Literature 20

1. Introduction

1.1Background and Context

Corporate sustainability reporting has expanded rapidly over the past decade, driven by mounting investor demand for Environmental, Social, and Governance (ESG) information, tightening regulatory requirements, and growing public scrutiny of corporate climate claims. Firms now communicate their environmental performance through an increasingly heterogeneous mix of narrative disclosures, standardised ESG scores, voluntary commitments such as net-zero targets and Science Based Targets initiative (SBTi) pledges, and visually elaborate sustainability reports. At the same time, concerns have grown that this expansion in disclosure volume has not been matched by a corresponding improvement in the veracity of what is disclosed. "Greenwashing" — broadly, the practice of overstating or misrepresenting environmental performance has become a central concern for regulators, investors, and civil society alike, precisely because the sophistication of corporate communication has begun to outpace the capacity of manual auditing and third-party assurance processes to verify it (Sulkowski, 2025).

This gap between the scale of ESG disclosure and the capacity for independent verification has motivated a rapidly growing body of research applying artificial intelligence (AI) and machine learning (ML) to the problem of greenwashing detection. Researchers have approached this problem from several distinct methodological traditions: natural language processing (NLP) techniques that analyse the text of sustainability reports for rhetorical inflation and vague or unsubstantiated claims; econometric and structural models that quantify the discrepancy between a firm's disclosed ESG performance and its verifiable operational outcomes; and, more recently, multimodal approaches that attempt to audit the visual and design elements of sustainability reports directly. Each of these traditions offers a partial view of the same underlying problem — the widening space between what firms say about their environmental performance and what they do.

1.2 Importance of the Topic

Understanding and detecting greenwashing matters for reasons that extend well beyond academic interest. Economically, ESG-linked capital allocation now influences trillions of dollars in institutional investment, meaning that undetected greenwashing can materially misdirect capital toward firms whose environmental claims are not supported by their underlying operations. Regulatory frameworks such as the EU's Corporate Sustainability Reporting Directive (CSRD) and European Sustainability Reporting Standards (ESRS) have made climate disclosure a matter of legal compliance rather than voluntary communication, raising the stakes of inaccurate or misleading reporting considerably (Al-Zoubi et al., 2026). At the same time, the growing use of AI systems to both generate and audit sustainability disclosures introduces a further governance dimension: as corporations increasingly rely on generative AI to draft ESG narratives, and as regulators increasingly rely on AI based tools to screen those narratives, the reliability, consistency, and legal defensibility of these automated systems becomes a pressing practical concern (Singh & Singh, 2024). The question of how to

detect greenwashing algorithmically is therefore not merely a technical exercise but one with direct consequences for capital markets, environmental accountability, and corporate legal liability.

1.3 Review of Existing Literature

Existing research on AI-driven greenwashing detection can be grouped into three broad methodological strands. The first strand comprises NLP and transformer-based approaches, which have evolved from early lexical and keyword-based methods through dense transformer models (BERT, RoBERTa, ClimateGPT) to more recent Aspect-Action Analysis (AAA) frameworks that explicitly distinguish verified historical actions from vague aspirational language (Ong et al., 2025; Braun et al., 2026). This strand has demonstrated that while topic classification of climate-related text is now highly accurate, assessing whether a given claim reflects genuine commitment or "cheap talk" remains difficult even for human experts, as reflected in the low inter-annotator agreement reported on benchmarks such as ClimateEval (Kurfalı et al., 2025).

The second strand consists of quantitative, structural models that operationalise greenwashing through the Disclosure-Performance Gap (DPG), or "Talk-Walk Gap" — the measurable discrepancy between a firm's narrative ESG positioning and its verified operational outcomes, typically emissions trajectories (Zeng, Wang & Zeng, 2025; Li et al., 2025). This literature has produced ensemble models combining optimisation algorithms, gradient boosting, and SHAP interpretability to predict greenwashing risk from financial and emissions data. However, this strand has also revealed a significant vulnerability: the magnitude of the measured Talk-Walk gap depends heavily on which ESG rating framework is used to proxy "talk," a phenomenon termed the "rating lens" problem (Kumar & Sieradzki, 2026), and a related blind spot to "crosswashing," whereby firms report truthful but peripheral sustainability investments to obscure the environmental impact of their core operations (Hassani & Bahini, 2024).

The third strand addresses multimodal auditing, in which commercial multimodal large language models (MLLMs) are used to evaluate the visual and design elements of sustainability reports. This literature has surfaced a counterintuitive and troubling finding — the "greenwashing paradox" — in which high-performing, professionally designed sustainability reports are more likely to be flagged as deceptive by commercial MLLMs than sparse, poorly designed reports from weaker environmental performers (Jakubczak et al., 2026).

Compounding this problem, inter-rater agreement across different commercial MLLMs auditing the same disclosures has been found to be close to zero, indicating that these models apply inconsistent and largely arbitrary evaluative heuristics.

Running through all three strands is a parallel governance literature, which argues that AI-generated or AI-audited ESG disclosures raise unresolved legal questions of accountability, given that automated systems can obscure human decision-making within a "volitional vacuum" that complicates traditional standards of corporate liability (Singh & Singh, 2024; Sulkowski, 2025).

1.4 Research Gap

Despite the depth of work within each of these three methodological strands, several important gaps remain, and it is at their intersection that the present study is positioned.

First, there is a methodological gap: research has developed largely in isolation across text-only NLP models, quantitative structural DPG models, and multimodal visual auditing models, with comparatively little work integrating these dimensions into a single verification framework capable of comparing narrative claims against both financial/operational evidence and visual presentation simultaneously (Zeng et al., 2025; Calamai et al., 2025).

Second, there is an empirical gap concerning the stability of greenwashing as a measurable construct. The rating lens dependency documented by Kumar and Sieradzki (2026) demonstrates that measured associations between disclosure and performance can shift dramatically depending on which "talk" and "walk" proxies are chosen, yet few studies have systematically tested how sensitive a given greenwashing-detection model is to these proxy choices using an independently assembled dataset.

Third, there is a theoretical and practical gap concerning the relationship between disclosure infrastructure and detected greenwashing prevalence. It remains unclear whether regulatory environments with more mature disclosure infrastructure (such as the EU's CSRD/ESRS regime) produce genuinely lower rates of greenwashing, or whether they simply make existing disclosure-performance gaps more visible and detectable — a distinction with significant implications for how cross-jurisdictional greenwashing statistics should be interpreted.

Finally, there is a geographical gap: the existing literature is concentrated on European and North American firms, with limited representation of major emerging-market disclosure regimes such as India's Business Responsibility and Sustainability Reporting (BRSR) framework, and correspondingly limited cross-jurisdictional evidence outside the US–EU axis.

1.5 Research Problem

Taken together, these gaps point to a specific and researchable problem: existing greenwashing-detection frameworks tend to treat "talk" and "walk" as measurable in a single, stable way, and tend to evaluate disclosure, financial performance, and visual presentation as separate rather than jointly calibrated signals. It remains unclear how robust a quantitative Disclosure-Performance Gap measure is to alternative operationalisations of firm performance, how such a measure behaves across sectors and regulatory jurisdictions, and what a verification architecture would need to look like in order to integrate narrative, financial, and (eventually) visual evidence into a single, defensible greenwashing-risk assessment rather than a standalone predictive label.

1.6 Research Questions and Objectives

This study addresses the following central research question:

To what extent can a quantitative Disclosure-Performance Gap framework, built from firm-level ESG and emissions data, reliably capture the divergence between corporate ESG "talk" and operational "walk," and what does the behaviour of this framework reveal about the

broader methodological challenges — rating-lens dependency, crosswashing, and cross-jurisdictional detection sensitivity — identified in the existing literature?

It is important to state at the outset that the empirical scope of this study is confined to the **quantitative, panel-data component** of the broader detection-and-mitigation pipeline outlined in the literature review. The study tests only the numerical Disclosure-Performance Gap between reported ESG scores and verified emissions outcomes; it does not process, and makes no empirical claims about, visual report design, narrative text, or iXBRL-tagged structured filings. The multimodal Dual-Branch Contrastive Calibration architecture and the Split iXBRL-Aware Agentic Framework remain **conceptual proposals**, grounded in the literature synthesis, that are put forward as the basis for a subsequent Phase 3 study rather than as components evaluated here.

With this scope in mind, the specific objectives of the study are to:

Examine the descriptive prevalence and cross-sectional distribution of greenwashing risk across sectors (Energy, Utilities, Industrials), countries, and time (2010–2024) within the assembled 450-firm-year panel dataset.

Operationalise and empirically compare alternative Disclosure-Performance Gap (DPG) measures using the available firm-level ESG and emissions variables, in order to assess the sensitivity of detected greenwashing risk to the choice of underlying performance proxy.

Test the "crosswashing" hypothesis that unverified sustainability commitments are more likely to be associated with elevated greenwashing risk than verified ones, using the verification-status variable available in the panel dataset.

Investigate whether cross-jurisdictional differences in detected greenwashing rates (EU/EEA+ versus USA) observed in the dataset are better explained by differences in underlying corporate behaviour or by differences in disclosure-infrastructure maturity and detection sensitivity.

Evaluate the performance and, critically, the limitations of machine-learning classifiers (Logistic Regression, Random Forest, Gradient Boosting) trained on the panel dataset to predict greenwashing risk from firm-level financial, emissions, and disclosure variables, with particular attention to the risk of target leakage arising from the DPG construction itself.

Situate these empirical findings — which are limited to the quantitative DPG component of the pipeline — within the broader literature on AI-driven greenwashing detection, and use them to inform, at a conceptual level, the design of a proposed Phase 3 Dual-Branch Contrastive Calibration architecture and Split iXBRL-Aware Agentic Framework capable of jointly evaluating narrative disclosure, visual presentation, and verified operational performance. This sixth objective is explicitly developmental and literature-based rather than empirical: no visual, narrative, or iXBRL data are analysed in the present study.

1.7 Significance of the Study

This study contributes to the literature in several ways. Empirically, it provides an independent, firm-level test of the rating-lens dependency and crosswashing hypotheses previously identified in the literature, using a dataset distinct in sectoral composition (Energy, Utilities, Industrials) and time frame (2010–2024) from prior studies. Methodologically, it offers a transparent worked example of how a Disclosure-Performance Gap index can be constructed, interpreted, and stress-tested for circularity and target leakage — a methodological caution that

is often acknowledged in principle but rarely demonstrated empirically in the applied ESG-analytics literature. Practically, the findings are intended to inform the design of the proposed Phase 3 Dual-Branch Contrastive Calibration model, by identifying precisely which quantitative signals (disclosure scores, emissions trajectories, verification status, jurisdiction) are robust enough to serve as an operational-reality anchor against which narrative and visual ESG communication can eventually be calibrated. In this way, the study seeks to move the broader research agenda one step closer to an integrated, verifiable, and legally defensible framework for greenwashing detection — one capable of distinguishing genuine disclosure-performance misalignment from the artefacts of measurement choice, detection sensitivity, or dataset construction.

2. LITERATURE REVIEW

2.1 Executive Summary

Artificial intelligence is now integral to corporate sustainability reporting, allowing algorithmic systems to scrutinize greenwashing across vast and varied disclosures (Calamai et al., 2025; Rosak-Szyrocka, 2026). Companies communicate environmental performance through narrative text, audited financials, structured KPIs, and visually persuasive documentation. This volume and structural diversity exceed the capacity of manual auditing to verify (Sulkowski, 2025). Researchers have proposed NLP, predictive ML, and multimodal AI as automated detection tools (Vinella et al., 2024; Braun et al., 2026), yet significant fractures persist. Greenwashing's definition remains legally underdetermined, ESG rating frameworks diverge empirically, models lack domain-specific context sensitivity, and multimodal systems frequently misattribute visual elegance to deceptive intent (Kumar & Sieradzki, 2026; Jakubczak et al., 2026). This review synthesizes the textual, structural, multimodal, and governance literature and proposes an integrated, agentic system for automated greenwashing detection within mandatory reporting environments.

2.2 NLP and Transformer-Based Frameworks

Early lexical and dictionary methods like TF-IDF, SVM, and Naive Bayes classify climate terminology effectively but falter on strategic ambiguity and rhetorical inflation (Kurfalı et al., 2025). On the ClimateEval benchmark, topic classification approaches near-perfect consensus while judging genuine commitment versus "cheap talk" yields a Krippendorff's alpha of just 0.1703 (Ong et al., 2025). This evidence shows that even human experts disagree on what constitutes greenwashing. Aspect-Action Analysis (AAA) improves on this by mapping environmental aspects to corporate actions along an ordinal commitment hierarchy: historical/verifiable, future/auditable, or vague/aspirational (Ong et al., 2025). The COGLM model extends AAA with LoRA-based contrastive-and-ordinal fine-tuning to resist overfitting across industries (Braun et al., 2026). Notably, domainspecific pretraining is not obviously necessary. ClimateGPT's 4.2-billion-token climate pretraining yielded only a 0.05 macro-F1 gain over base Llama-2 (Kurfalı et al., 2025), while lightweight PEFT-tuned open models (Llama-3-8B, Gemma-7B) match or exceed GPT-4o on sentence-level detection (Birti et al.,

2025; Florstedt et al., 2025). A further limitation, the "paradox of context" (Schmoll & Jatowt, 2025), shows that feeding LLMs raw, boilerplate-heavy PDFs collapses zero-shot accuracy, while pre-extracted structured metadata restores it.
Agentic retrieval must precede semantic analysis.

2.3 Quantitative Modeling and the Disclosure-Performance Gap (DPG)

Beyond text, greenwashing manifests financially. The DPG, or "Talk-Walk Gap," measures the discrepancy between disclosure and verified outcomes. Ensemble models combining Improved Hunter-Prey Optimization, XGBoost, and SHAP predict this gap ahead of public release, with SHAP identifying firm size, shareholder concentration, ROA/ROE, and asset turnover as key predictors (Zeng et al., 2025). Random-Forest "Pressure Pool" models link financial stress, regulatory scrutiny, and media attention to manipulation (Li et al., 2025). These models are ontologically unstable. Across 200 carbon-intensive European firms, the Talk-Walk gap's magnitude swings from r = 0.68 (using CDP scores) to r = 0.12 (using LSEG/Refinitiv scores) for the same firms (Kumar & Sieradzki, 2026). This shows greenwashing is measurement-dependent, not a fixed attribute. A related blind spot is "crosswashing" (Hassani & Bahini, 2024): firms accurately report peripheral sustainability investments (e.g., an oil major's community solar projects) while obscuring core-business harm. Each fact is truthful, yet the aggregate impression misleads, evading detection by aggregated financial or surface-text models alone.

2.4 Multimodal Auditing and the Greenwashing Paradox

Reports increasingly use visual rhetoric, nature photography, dashboards, and bespoke palettes to build legitimacy, prompting researchers to test commercial MLLMs (GPT-4o/5.1, Claude 3.5/4.5, Gemini 2.5) as visual auditors (Jakubczak et al., 2026). The results are counterintuitive. Rather than the negative correlation signaling theory predicts between verified performance and assigned risk, ChatGPT shows a statistically significant positive correlation ($r = 0.50$, $p = 0.025$). High performers who invest in polished reporting are accused of greenwashing more, while poor performers with sparse reports evade detection (Jakubczak et al., 2026). This compounds with near-zero inter-rater reliability across models auditing identical disclosures (Krippendorff's $\alpha = 0.021$) and documented penalization of exemplary human-rights records simply for featuring polished photography. Zeroshot commercial generative AI is unsuited to regulatory assurance without calibrated, human-in-the-loop oversight (Jakubczak et al., 2026).

2.5 Legal Liability, Algorithmic Accountability, and Governance

AI-driven reporting raises acute governance dilemmas. The "Garbage In, Garbage Out" and the "Second-Person Trust" problem (Sulkowski, 2025) highlight that AI can validate that data has not been altered post-ingestion (Third-Person Trust) but not that it was accurately measured at its physical source (Second-Person Trust). This vulnerability is acute in Scope 3 and supply-chain reporting, where automation can become a "legitimacy shield." Autonomous or semi-autonomous generation of misleading disclosures also complicates traditional means rebased fraud liability, creating a corporate "volitional vacuum" in which firms disclaim culpability by

blaming "black-box" algorithmic error (Singh & Singh, 2024; Baghel & Desai, 2025). Legal scholars argue courts across the US, EU, and India should reject this defense (Al-Zoubi et al., 2026). The regulatory response, exemplified by the EU's CSRD/ESRS, converges on three non-negotiable mechanisms: mandated algorithmic due diligence (disclosed training data, weights, bias mitigation), enforceable human oversight (non-delegable board sign-off), and multi-tier physical verification (IoT emissions monitoring, satellite telemetry, certified ledgers) (Al-Zoubi et al., 2026; Bergougui et al., 2026; Sulkowski, 2025).

2.6 Research Gaps and the Proposed Contribution

The literature converges on four structural gaps preventing operational deployment, mapped below against this study's proposed architectural response: a Split iXBRL-Aware Agentic Verification Framework. Leveraging the EU's mandatory European Single Electronic Format (which pairs human-readable XHTML with machine readable iXBRL tags), a High-Recall Retrieval Agent parses structured filings directly, extracting KPI and CapEx/OpEx nodes while discarding aesthetic boilerplate. A Focused Verification Agent then performs quantitative KPI regression, aligning narrative commitments against audited financial flows under EU Taxonomy criteria. A Dual-Branch Contrastive Calibration layer directly targets the greenwashing paradox: a Visual Rhetoric Encoder (Branch A) scores design sophistication while an Operational Performance Encoder (Branch B) encodes verified Scope 1–3 and CapEx transition data. Training under a joint contrastive loss penalizes visual polish only when it diverges from real operational performance, so aesthetic quality is never mistaken for guilt.

Table 1: Empirical Limitation in Current State-of-the-Art and Proposed Architectural/Methodological Contribution

Identified Research Gap	**Empirical Limitation in Current State-of-the-Art**	**Proposed Architectural / Methodological Contribution**
Context Degradation in PDFs ("Paradox of Context")	Zero-shot LLMs suffer severe performance degradation when digesting unstructured 100+ page corporate PDFs (Schmoll & Jatowt, 2025).	Split iXBRL-Aware Agentic Workflow: isolates machine-readable XML/HTML metadata tags, bypassing visual noise and formatting clutter.
Multimodal "Greenwashing Paradox"	Commercial MLLMs exhibit a positive correlation (r = 0.50) with greenwash accusations, penalizing high performers for graphic design (Jakubczak et al., 2026).	Dual-Branch Contrastive Calibration: separates visual-aesthetic encoding from operational-emissions grounding, penalizing style only when uncoupled from data.

Identified Research Gap	Empirical Limitation in Current State-of-the-Art	Proposed Architectural / Methodological Contribution
Text–Performance Schism & Rating Divergence	Text models ignore financial balance sheets; quantitative DPG models diverge based on arbitrary commercial rating lenses (Kumar & Sieradzki, 2026).	Activity-Level Financial Grounding: correlates textual promises directly with audited EU Taxonomy CapEx/OpEx investments rather than commercial ratings.
Algorithmic "Volitional Vacuum" & Legal Risk	Firms disclaim liability for AI-generated hallucinations and deceptive claims; no audit trail exists (Singh & Singh, 2024; Sulkowski, 2025).	Verifiable Audit Trail & Agentic Logs: structured reasoning chains, citation of precise iXBRL financial nodes, human- in-the-loop governance.
Multimodal "Greenwashing Paradox"	Commercial MLLMs exhibit a positive correlation (r = 0.50) with greenwash accusations, penalizing high performers	Dual-Branch Contrastive Calibration: separates visual-aesthetic encoding from operational-emissions grounding,

2.7 Conclusion

Detection has advanced from keyword counts to transformers, Aspect-Action Analysis, and multimodal vision language models, exposing methodological, empirical, and governance vulnerabilities at each step. These include low inter-annotator agreement, rating-lens-dependent DPG instability, the multimodal greenwashing paradox, and PDF-induced context degradation. Theoretically, greenwashing must be evaluated as a multidimensional misalignment between narrative, visual signaling, and audited operations, not a purely linguistic anomaly. Practically, off-the-shelf generative AI cannot serve as an autonomous compliance auditor. Mandatory regimes like CSRD/ESRS require verifiable architectures that close the "volitional vacuum" through transparent provenance and algorithmic accountability. Future research must prioritize standardized, legally validated ground truth benchmarks and activity-level crosswashing detectors, so computational sustainability tools foster genuine decarbonization rather than more sophisticated algorithmic deception.

3. METHODOLOGY

3.1 Research Design

This study adopts a Design Science Research (DSR) design, following the process model of Peffers, Tuunanen, Rothenberger, and Chatterjee (2007). DSR is appropriate because the overarching research objective is not merely to describe or explain the phenomenon of ESG greenwashing, but to identify, justify, and specify the requirements for a novel artefact — a

Split iXBRL-Aware Agentic Verification Framework incorporating a Dual-Branch Contrastive Calibration layer — capable of detecting the discrepancy between corporate disclosure ("Talk") and audited operational performance ("Walk").

Peffers et al. (2007) specify six sequential activities: (1) problem identification and motivation, (2) definition of the objectives for a solution, (3) design and development, (4) demonstration, (5) evaluation, and (6) communication. The work reported in this paper corresponds to Activities 1 and 2, together with an anticipatory (ex-ante) component of Activity 3. Specifically:

Activity 1 (Problem Identification and Motivation) was addressed through a structured synthesis of the textual, quantitative, multimodal, and governance literature on AI-driven greenwashing detection, from which four structural research gaps were derived (context degradation in unstructured filings, the multimodal "greenwashing paradox," text–performance schism, and the algorithmic "volitional vacuum").

Activity 2 (Definition of the Objectives for a Solution) was informed by a secondary statistical analysis of a firm-level ESG panel dataset, undertaken to empirically test — rather than assume — the motivating claims of the literature (e.g., the existence of a Talk–Walk gap, the rating-lens dependency of greenwashing measures, and the feasibility of purely disclosure-based prediction).

This dataset analysis functions as a diagnostic, ex-ante evaluation that precedes and informs — rather than validates — the artefact itself. It establishes the empirical requirements that the artefact must satisfy (notably, that a disclosure-only branch is insufficient and must be paired with an independently verified operational-evidence branch). The design and development of the artefact (Activity 3, in full), its demonstration (Activity 4), and its formal evaluation (Activity 5) are reserved for the subsequent phase of the project ("Phase 3") and fall outside the scope of the present paper.

The study is therefore best characterised, in this phase, as problem-centered and diagnostic-empirical DSR, rather than artefact-evaluative DSR.

3.2 Nature of Research

The research is mixed-methods and empirical, combining a qualitative documentary synthesis with quantitative secondary-data analysis, both oriented toward the specification of a conceptual/architectural design rather than the collection of primary data. The literature synthesis is qualitative and interpretive, drawing on published theoretical, empirical, and doctrinal/regulatory sources. The dataset analysis is quantitative and empirical, applying descriptive statistics, cluster-robust significance testing, and supervised machine-learning classification to a secondary firm-year panel. No primary data were collected by the research team; all evidence is drawn from previously published academic literature and previously compiled secondary datasets.

3.3 Sample & its Technique

3.3.1 Literature Sample

The unit of analysis for the literature review is the individual published work (journal article, conference paper, or preprint). The corpus was purposively (non-probabilistically) selected using keyword combinations pairing "greenwashing" with "AI," "machine learning," "NLP," "multimodal," "ESG rating," and "algorithmic accountability," searched across arXiv, SSRN, and Scopus/Web of Science-indexed journal databases. Inclusion criteria required that a source (i) directly address computational or AI/ML-based detection of corporate greenwashing, ESG rating divergence, or the legal/regulatory governance of AI-generated sustainability disclosures, and (ii) be published between 2024 and 2026, reflecting the recency of methodological development in this sub-field; foundational works cited for methodological rather than substantive purposes (e.g., Peffers et al., 2007) were exempted from the recency criterion. The resulting corpus comprises twenty-two sources spanning peer-reviewed journals (e.g., Sustainability, PLOS ONE, Scientific Reports, Journal of Business Ethics), peer reviewed conference proceedings (ACL 2025, ICAIF '25), and preprint repositories (arXiv, SSRN), and is documented in full in the accompanying reference list.

3.3.2 Dataset Sample

The unit of analysis for the statistical component is the firm-year observation. The panel comprises 30 publicly listed firms drawn in equal number (n = 10 each) from three carbon-intensive sectors identified as high-risk in the literature — Energy, Utilities, and Industrials — observed annually from 2010 to 2024, yielding 450 firm-year observations. Firms were selected purposively, not randomly, subject to two explicit inclusion criteria: (i) sectoral membership in one of the three designated categories, and (ii) the availability of a complete set of financial, Scope 1–3 emissions, disclosure/commitment, and ESG-score variables across the full panel window, together with a pre-existing binary greenwashing_flag variable used as the outcome proxy (65 flagged firm-years; 14.4% of the panel). Geographic coverage is uneven by design of the underlying secondary dataset rather than by researcher selection, and is treated as a sampling limitation (Section 7): 17 of the 30 firms are US-domiciled (255 firm-years), 4 are German, 2 each are Italian and British, and the remaining firms (Norway, France, Switzerland, Spain, Luxembourg) are represented by a single firm each. Consistent with the guidance against mislabelling nonprobability sampling as random, this sample is explicitly identified as a purposive, criterion-based sample and not a representative random sample of listed firms.

3.3.3 Data Collection Tool

No primary instruments (questionnaires, interviews, or experiments) were used. Two secondary data-collection procedures were followed.

Documentary/literature sources: peer-reviewed articles, conference papers, and preprints were retrieved using computational literature-search tools and cross-verified against open-access academic repositories and primary regulatory texts (notably Directive (EU) 2022/2464, which established the Corporate Sustainability Reporting Directive (CSRD) and its associated

European Sustainability Reporting Standards (ESRS), and the EU Sustainable Finance Disclosure Regulation (SFDR)).

Secondary panel dataset: financial, emissions, and disclosure variables, together with the pre-existing greenwashing flag, were compiled from official and publicly available reporting and ratings platforms rather than collected by the research team. Thirty first-year-of-panel observations had no previous-year emissions observation, so their year-on-year emissions change could not be calculated. These undefined values were set to 0.0 solely as an explicit modelling convention to provide a numeric placeholder; this value does not mean that emissions did not change. Categorical commitment and CDP-grade variables were coded into numeric/ordinal scales prior to analysis.

Statistical processing comprised (i) descriptive and cluster-robust inferential testing (company-clustered standard errors, to correct for the panel's repeated-measures structure, in which 450 rows represent 30 firms observed repeatedly rather than 450 independent cases), and (ii) supervised classification using Logistic Regression, Random Forest, and Gradient Boosting classifiers evaluated by 5-fold cross-validation, with a leakage-controlled re-specification used to obtain an honest, unseen-firm estimate of predictive validity.

3.3.4 Variables Used

The dependent variable (Section 3 analysis) is the pre-existing binary greenwashing flag (1 = flagged, 0 = not flagged). The principal independent and derived variables, together with their operational definitions, are specified below.

Table 2: Operational Definitions of Independent and Derived Variables

Variable	Operational definition
Standardised score, z(x)	$z(x_{it}) = (x_{it} - \mu_x) / \sigma_x$, where μ_x and σ_x are the panel mean and standard deviation of variable x. Applied to ESG score, emissions change, and carbon intensity prior to index construction.
Year-on-year Scope 1 emissions change, ΔE	$\Delta E_{it} = [(Emissions_{it} - Emissions_{i,t-1}) / Emissions_{i,t-1}] \times 100$, expressed as a percentage. For the first panel year, the change is undefined because no previous-year observation is available; these cases were assigned 0.0 solely as a modelling placeholder and should not be interpreted as zero emissions change.
Disclosure–Performance Gap index, DPG (primary variant)	$DPG_{it} = z(ESGscore_{it}) + z(\Delta E_{it})$. Positive values indicate rising standardised disclosure scores concurrent with rising (i.e., worsening) emissions — the operational definition of the "Talk–Walk" discrepancy.
DPG, carbon-intensity variant	$DPG_{it}(CI) = z(ESGscore_{it}) + z(CarbonIntensity_{it})$, substituting standardised carbon intensity for emissions change as the operational ("Walk") proxy, used to test rating-lens/proxy sensitivity.

Variable	Operational definition
Greenwashing prevalence rate, GWR (referred to descriptively as the group greenwashing rate)	GWR_g = (Σ flagged firm-years in group g / Σ total firm-years in group g) × 100, computed by sector, country, and jurisdiction (EU/EEA+ vs. United States).
Predictive-model performance	Standard classification metrics — area under the receiver-operating-characteristic curve (AUC-ROC), F1-score, and recall for the flagged (positive) class — computed under two feature specifications: (a) an unrestricted specification including DPG, ESG score, and ΔE; and (b) a leakage-controlled specification from which these label-defining variables are excluded, used to obtain a genuine unseen-firm estimate.
Firm size	Operationalised as firm revenue (continuous), entered as a coefficient in the logistic-regression specification and as a feature-importance score in the Random Forest / Gradient Boosting ensembles.

Source: ESG Greenwashing Panel Dataset – Final Analysis; compiled from official and publicly available reporting and ratings platforms.

Control/contextual variables include sector (Energy, Utilities, Industrials), country and jurisdiction (EU/EEA+ vs. United States), net-zero/SBTi commitment status (verified vs. unverified), and CDP disclosure grade (ordinal).
All ratios and rates above are reported as descriptive statistics; inferential comparisons across groups (sector, country, jurisdiction) use company-clustered significance tests rather than the naive (unclustered) chi-square tests used in the
original, pre-correction analysis, in order to account for the non-independence of repeated firm-year observations.

3.4 Ethical Considerations

The study involves no human participants, no primary data collection, and no identifiable personal data; consequently, no informed-consent, withdrawal, or anonymisation procedures were required, and institutional ethics review was not applicable. Two study-specific safeguards were nonetheless observed. First, because the greenwashing flag in the secondary panel is a pre-existing modelling proxy rather than an independently reverified determination against primary regulatory sources (e.g., CDP Open Data, SEC EDGAR filings), firm-level results are reported and interpreted as a screening-level statistical association rather than as a finding of wrongdoing by any named entity, consistent with the responsible-use principle for public company-level data. Second, in accordance with responsible academic practice, all

literature and data sources are cited to their original authors and providers, and the use of computational literature-search and AI-assisted drafting tools in preparing the review was disclosed, with all analytical synthesis and conclusions critically evaluated and finalised by the research team.

3.5 Limitations

Small firm count. The panel comprises only 30 unique companies; firm-level findings (for example, the negative coefficient obtained for firm revenue) should not be generalised beyond this sample and, notably, run contrary to the direction reported by Zeng, Wang, and Zeng (2025) on a comparable Chinese-listed firm sample.

Non-probability sampling and uneven geographic coverage. Firms were purposively selected on data availability grounds rather than randomly sampled; 17 of 30 firms are US-domiciled, while several countries (Norway, France, Switzerland, Spain, Luxembourg) are represented by a single firm each, so country-level rates should be read descriptively rather than inferentially.

Panel (repeated-measures) structure. The 450 firm-year rows represent 30 firms observed repeatedly rather than 450 independent cases; although this is now corrected for using company-clustered testing, the correction narrows several previously reported significance levels (e.g., the EU-versus-US comparison, $p < 0.001$ uncorrected versus $p = 0.009$ corrected).

Label provenance. The greenwashing flag is not independently re-verified against primary sources and should be treated as a modelling proxy rather than confirmed ground truth.

Rating-lens (proxy) dependency. DPG estimates vary substantially depending on the operational performance proxy chosen ($r = 0.103$ using carbon intensity versus $r = 0.480$ using emissions trend in this sample), consistent with Kumar and Sieradzki (2026); this indicates that findings are, in part, an artefact of measurement choice rather than a fixed property of the firms studied.

Coverage gaps relative to the proposed artefact. The dataset contains no Indian firms (precluding the planned BRSR comparison), no iXBRL-tagged CapEx/OpEx data (precluding verification of claims against actual capital allocation), and no visual/report-design data (precluding testing of Phase 3's Dual-Branch aesthetic-calibration component); these gaps restrict what the present analysis can establish and are treated as requirements for the data to be assembled in Phase 3, not as findings.

Predictive validity. Once definitional-leakage and seen-firm evaluation are corrected for, genuine predictive power on unseen firms is weak (AUC = 0.559, recall = 6.7%); conclusions regarding the feasibility of automated prediction from disclosure-only variables are therefore provisional and motivate, rather than establish, the need for the independently verified operational-evidence branch proposed for Phase 3.

4. DATA ANALYSIS AND INTERPRETATION

4.1 Executive Summary

This report consolidates the original Phase 2 statistical analysis of the ESG Greenwashing Panel Dataset with the corrections applied after peer review. The dataset's most robust finding, flagged firms report higher ESG scores while simultaneously showing rising, not falling, emissions, holds up under every correction and remains the strongest evidence for a genuine Talk–Walk gap in this sample. The original report's headline predictive claim (97% accuracy, AUC = 1.000), however, did not survive scrutiny: it was driven by the model being given variables that help define its own target label, evaluated on firms already partly seen in training. Corrected for both issues, genuine predictive power on entirely unseen firms is weak (AUC = 0.559). Three of four significance tests held up under proper company-clustered statistical testing; one (EU vs. US) survived but was considerably overstated by the original naive test. This document explains what changed, why, and what it means for the project's next phase.

4.2 Dataset and Objective

The panel covers 30 firms (10 each in Energy, Utilities, and Industrials) from 2010–2024, with financial, emissions, and disclosure variables plus a pre-existing binary greenwashing_flag (65 flagged, 14.4%). The objective is to test whether a measurable gap between reported ESG performance (“Talk”) and operational emissions trends (“Walk”) can serve as a statistical greenwashing-risk signal. Missing year-on-year emissions values (30 observations, all first-year-of-panel cases) were set to 0.0 as a modelling convention; categorical commitment and CDP-grade variables were converted to numeric/ordinal scales for modelling.

4.3 Greenwashing Prevalence and the Talk–Walk Discrepancy

Prevalence differs descriptively by sector (Utilities 17.3%, Energy 15.3%, Industrials 10.7%) but this difference is not statistically significant once company-clustered testing replaces the naive chi-square test (p=0.40–0.78 by sector). Country-level rates vary widely, but per-country samples are very small (15–60 firm-years from 1–4 firms) and should be read as descriptive only - Italy's 43.3% rate, for instance, comes from just 2 firms over 15 years.

Table 3: Country-Level Greenwashing Prevalence and Sample Distribution

Country	GW Rate	N (firm-years)	Firms
Italy	43.3%	30	2
Norway / France	33.3%	15 each	1 each
UK	20.0%	30	2
Germany	16.7%	60	4
Switzerland	13.3%	15	1
USA	9.4%	255	17
Spain / Luxembourg	0.0%	15 each	1 each

The clearest and most robust finding in the dataset is the Talk–Walk discrepancy: flagged firm-years show materially higher ESG scores (mean 64.4 vs. 55.6 for non-flagged) while simultaneously showing rising rather than falling emissions (+2.83% vs. −3.82% year-on-year Scope 1 change) - a pattern that survives every correction applied and is not an artifact of the modelling issues found elsewhere. The Disclosure-Performance Gap (DPG) index (standardised ESG score + standardised YoY emissions change) captures this cleanly, rising from a corrected mean of −0.239 (non-flagged) to +1.414 (flagged).

Rating-lens dependency confirmed. Two DPG variants built from the same ESG score but different performance proxies correlate with the flag very differently - r=0.103 using carbon intensity vs. r=0.480 using emissions trend - directly replicating the literature's own finding that greenwashing measures are highly sensitive to which performance benchmark is chosen (reported elsewhere as r=0.12 for Refinitiv-style vs. r=0.68 for CDP-style ratings). This means the choice of "Walk" proxy, not just the detection method, materially changes whether greenwashing is "found."

4.4 Crosswashing, Net-Zero Paradox, and Predictive Modelling

Crosswashing hypothesis - not supported. Among firms with net-zero or SBTi commitments, verified firms actually show a higher flag rate (18.7%) than unverified firms (11.1%) - the opposite of what the "talk without walk" hypothesis predicts. This reversal remains non-significant even under the corrected clustered test (p=0.091), so it should be read as a genuine null result, not a contradicted-but-still-real effect. A related paradox: firms with net-zero targets show a slightly higher flag rate (15.9%) than firms without (12.2%), plausibly because prominent commitments generate more disclosure and therefore more opportunity for a measurable gap to appear - not evidence that setting targets causes greenwashing.

Predictive modelling - the corrected headline result. This is the most consequential revision in the analysis.

The original report benchmarked three classifiers by 5-fold cross-validation and reported the following:

Table 4: Original Predictive Model Performance (5-Fold Cross-Validation with Definitional Leakage)

Model (Original Report, 5-fold CV)	Features Used	ROC-AUC	F1 (GW)	Recall (GW)
Logistic Regression	All variables incl. DPG	0.903	0.609	82%
Random Forest	All variables incl. DPG	0.994	0.916	92%
Gradient Boosting	All variables incl. DPG	0.992	0.928	89%

These figures looked strong, but all three models were given the same variables used to help define the greenwashing label itself. Removing that leakage and testing on companies genuinely absent from training produces a very different picture:

Table 5: Comparison of Original Headline Results and Corrected Unseen-Firm Test Performance

Model / Test Set	Features	AUC	Recall (GW)	Interpretation
Original headline result	All variables incl. DPG, ESG score, YoY emissions	1.000	1.00	Inflated by definitional leakage — not a genuine finding
Corrected: unseen-firm test	Leakage-controlled (DPG, ESG score, YoY emissions removed)	0.559	0.067	Honest estimate: weak real-world signal from independent variables alone

The originally reported near-perfect accuracy reflected the model being given variables that help define its own target, evaluated on firms it had already partly seen. Once both issues are removed, real-world predictive power from independent financial, emissions-level, and governance variables alone is weak (AUC 0.559, 6.7% recall) - the dataset does not yet support a reliable standalone predictive model for firms outside the training sample. The strongest surviving predictors under the corrected, leakage-controlled feature set are Scope 2/3 emissions levels, carbon intensity, and firm revenue, not verification status or net-zero commitments.

Firm size: a genuinely unresolved question. The corrected model's evidence on firm size directly illustrates why single-dataset findings should be treated cautiously in this literature:

Table 6: Firm Size Impact: Empirical Findings and Comparison to Prior Literature

Evidence (this dataset)	Direction Found	Comparison to Prior Studies
Logistic regression coefficient on firm revenue	−0.31 (negative: larger firms → lower flag risk)	Aligns with Zhang & Qi (2026); contradicts Zeng, Wang & Zeng (2025), who found the opposite direction on a comparable Chinese-listed-firm sample

Evidence (this dataset)	Direction Found	Comparison to Prior Studies
Ensemble (RF/GB) feature importance for revenue	5.5–7.4% (moderate, not dominant)	Firm size matters somewhat but is far from the top predictor in either direction

4.5 Cross-Jurisdictional Findings and Data Limitations

EU/EEA+ firms show a markedly higher flag rate than US firms (21.0% vs. 9.4%); this remains significant after clustering ($p=0.009$) but with materially less confidence than the original naive test suggested ($p<0.001$). Two explanations remain open: stricter EU disclosure regimes (CSRD, SFDR) may make greenwashing more detectable rather than more prevalent, or the EU sample may simply skew toward larger, more visible reporters.

Label provenance: the flag is not independently re-verified against primary sources (CDP Open Data, SEC EDGAR); it should be treated as a given proxy for modelling, not confirmed ground truth.

Small firm count: only 30 unique companies populate the panel; firm-level findings (see the firm-size comparison above) should not be generalized beyond this sample.

Panel structure: 450 rows are 30 firms measured repeatedly, not 450 independent cases; this is now corrected for in significance testing (Section 2) but was not part of the original modelling pipeline.

Coverage gaps: no Indian firms (the methodology's planned BRSR comparison cannot be tested); no iXBRL-tagged CapEx/OpEx data (claims cannot yet be checked against actual capital allocation); and no visual/report-design data (Phase 3's aesthetic-calibration branch cannot be tested with this dataset alone).

4.6 Implications for Phase 3 and Conclusion

The corrected results sharpen rather than undermine the case for Phase 3's Dual-Branch Contrastive Calibration design: a narrative/disclosure branch alone (of which ESG score and emissions self-reporting are proxies) is demonstrably insufficient to reliably flag greenwashing in unseen firms, reinforcing the need for the operational reality branch to rely on genuinely independent, audited evidence rather than further self-reported ESG indicators.

The recommended pipeline going into Phase 3:

1. Claim Detection - identify environmental/climate claims in corporate disclosures.

2. Semantic/Narrative Analysis - assess claim strength, specificity, sentiment, and visual presentation.
3. Operational Evidence Extraction - pull emissions, targets, and CapEx/OpEx indicators.
4. DPG Calculation - quantify the gap between ESG "Talk" and operational "Walk."
5. Dual-Branch Contrastive Calibration - compare narrative vs. operational representations.
6. Independent Verification - cross-check high-risk cases against regulatory filings and external data.
7. Risk Flag - output a screening score requiring human review, not a legal determination.

In summary: the Talk–Walk discrepancy and the rating-lens dependency are real, robust patterns in this data and should anchor Phase 3; the near-perfect predictive claim in the original analysis was a methodological artifact and has been replaced with an honest, modest estimate; and the crosswashing and net-zero paradox findings are credible null/reversed results, not confirmations, of the hypotheses they were designed to test. This is a stronger, more defensible foundation for Phase 3 than the original report provided, even though several of its headline numbers are now less impressive.

5. Discussion

5.1 The Central Talk–Walk Discrepancy

The most consistent finding across the panel is that flagged firm-years combine a materially higher mean ESG score (64.4 vs. 55.6) with a positive year-on-year change in Scope 1 emissions (+2.83% vs. −3.82%) relative to non-flagged observations. Interpreted substantively, this pattern indicates that greenwashing risk in this sample is not driven by weak disclosure or poor ESG positioning as such, but by a *divergence* between what a firm reports about itself and what its underlying emissions trajectory shows. In other words, the firms most likely to be flagged are not the worst ESG performers on paper — they are, if anything, the ones presenting the strongest narrative while their operational reality is moving in the opposite direction.
This result is directly consistent with the Disclosure-Performance Gap (DPG) or "Talk–Walk Gap" construct developed in the econometric literature (Zeng, Wang & Zeng, 2025; Li et al., 2025), where IHPO–XGBoost– SHAP ensembles were similarly able to isolate a gap between narrative disclosure and verified performance as the operative signal of manipulation risk, rather than treating either dimension in isolation. The correspondence between the two studies strengthens the interpretive claim that greenwashing, at least in structural quantitative terms, is best conceptualised relationally — as a discrepancy variable — rather than as a property of either the "talk" or the "walk" side of the ledger on its own. A plausible explanation for why this discrepancy arises is that firms under pressure to maintain investor and stakeholder

confidence may have stronger incentives to *report* ambitiously than to alter capital-intensive physical operations quickly; ESG narrative can be revised annually, while emissions trajectories are constrained by asset lifecycles, energy mix, and transition costs that cannot move at the same speed. This asymmetry between the cost of narrative adjustment and the cost of operational adjustment offers a substantive, non-speculative account of why the gap would be expected to open precisely among firms with elevated ESG scores.

5.2 The Rating-Lens Dependency Problem

A second and arguably more methodologically consequential finding is that the two DPG measures constructed from the same ESG score but different performance proxies produced markedly different correlations with the greenwashing flag: r = 0.103 for DPG-1 (ESG score vs. carbon intensity) against r = 0.480 for DPG-2 (ESG score vs. YoY Scope 1 change). This is not a minor discrepancy — it represents an almost fivefold difference in apparent predictive validity depending solely on which "performance" proxy is chosen while holding the "talk" side constant.

This finding replicates, in a new empirical setting, the "rating lens" effect documented by Kumar and Sieradzki (2026), who found that the Talk–Walk gap among 200 carbon-intensive European firms swung between r = 0.68 (CDP-based) and r = 0.12 (LSEG/Refinitiv-based) depending purely on the ESG rating framework used to operationalise "talk." The present analysis extends this observation to the performance side of the equation: even when the disclosure measure is held constant, the choice of *performance* proxy (a stock measure, carbon intensity, versus a flow measure, year-on-year emissions change) is sufficient to move the correlation from a range associated with weak discriminatory power (DPG-1) to one associated with moderate discriminatory power (DPG2). Taken together with Kumar and Sieradzki's findings, this suggests that greenwashing is not a fixed or objectively measurable firm attribute but a construct whose magnitude is highly sensitive to the specific measurement architecture chosen by the researcher or rating agency. Practically, this raises a serious concern for any AI/ML detection system that is trained on a single greenwashing label without disclosing or interrogating how that label's underlying "talk" and "walk" proxies were selected: the model may simply be learning the idiosyncrasies of the labelling convention rather than a stable, generalisable phenomenon.

5.3 The Crosswashing Hypothesis: An Unexpected, Reversed Finding

Contrary to the crosswashing hypothesis advanced by Hassani and Bahini (2024) — which predicts that unverified sustainability commitments should cluster with higher greenwashing risk — this dataset shows the opposite direction: among firms with net-zero or SBTi commitments, verified firms exhibited a *higher* greenwashing flag rate (18.7%) than unverified firms (11.1%), although the difference did not reach conventional statistical significance (χ^2 = 2.168, p = 0.141; Fisher's OR = 1.842, p = 0.121).

This reversal should not be dismissed as a null result; it is informative in its own right. One plausible explanation is a form of detection-mechanism circularity: if third-party verification processes are themselves the channel through which discrepancies between reported and operational performance become visible, then verified firms may simply be *more likely to be*

caught rather than more likely to be genuinely deceptive, while unverified firms' claims may go unchallenged for lack of an independent audit trail. A second explanation, consistent with the broader dataset limitations, is that the greenwashing_flag construction may not have been designed to test the crosswashing mechanism specifically, and with only 270 committed firm-years split 11/32 across the two verification categories, the analysis is statistically underpowered to detect an effect in either direction. This finding illustrates a broader point emphasised throughout the literature review: greenwashing detection frameworks that rely on a single binary label are vulnerable to conflating distinct underlying mechanisms — misrepresentation, non-disclosure, and crosswashing are conceptually separable phenomena (Hassani & Bahini, 2024) that a single flag cannot cleanly disentangle.

5.4 The Net-Zero Commitments Paradox

Firms with a stated net-zero target showed a somewhat higher greenwashing flag rate (15.9%) than firms without one (12.2%). Read alongside the Talk–Walk finding above, a coherent interpretation emerges: firms that make prominent, ambitious climate commitments generate a larger and more visible ESG "talk" surface, which mechanically increases the *opportunity* for a measurable disclosure-performance gap to open, even absent any difference in underlying intent. This is consistent with, rather than contradictory to, the DPG framing developed in Section 5.1 — the more ambitious the stated commitment, the more exposed a firm becomes to being flagged when its emissions trajectory does not keep pace. This should be read as a hypothesis about measurement exposure rather than a claim that net-zero commitments causally produce greenwashing behaviour; the dataset's descriptive, cross-sectional design cannot adjudicate between this "exposure" account and the alternative explanation that firms with ambitious targets are disproportionately drawn from harder-to-decarbonise, asset-heavy transition environments.

5.5 Cross-Jurisdictional Divergence: Detection Sensitivity over Prevalence

EU/EEA+ firms showed a substantially higher greenwashing rate (21.0%) than US firms (9.4%; $\chi^2 = 11.14$, $p = 0.0008$), a direction that runs counter to the naïve expectation that stronger disclosure infrastructure (EU Taxonomy, CSRD, SFDR) should suppress rather than elevate detected greenwashing. This finding connects meaningfully to the governance literature reviewed above. Sulkowski (2025) and Al-Zoubi et al. (2026) argue that mandatory, structured disclosure regimes such as CSRD/ESRS are designed precisely to make discrepancies between narrative and operational reality *more* visible and auditable, not to eliminate the underlying discrepancies themselves. Read this way, the higher EU rate is better understood as evidence of superior detection capacity within a stricter regulatory environment than as evidence that European firms are more prone to greenwashing behaviour in an absolute sense. An alternative, non-exclusive explanation is compositional: the EU/EEA+ subsample is dominated by large, high-profile ESG reporters for whom disclosure-performance gaps are simply more measurable and more publicly scrutinised. Both explanations point toward the same broader conclusion echoed throughout the literature review — that observed greenwashing "prevalence" is partly an artefact of a jurisdiction's disclosure-infrastructure maturity rather than a pure measure of underlying corporate behaviour.

5.6 Predictive Performance and the Limits of the Modelling Exercise

While the Random Forest and Gradient Boosting classifiers achieved very high discriminative performance (ROC-AUC of 0.994 and 0.992 respectively, rising to 1.0000 on the held-out test set), this result should be interpreted with considerable caution rather than as evidence of a generalisable detection capability. Because the DPG index and its constituent variables (esg_score_0_100, yoy_scope1_change_pct) appear both in the feature set and, conceptually, in the construction logic of the greenwashing_flag itself, the near-perfect performance is more parsimoniously explained by structural circularity than by genuine predictive signal. This mirrors a broader caution raised in the literature: Schmoll and Jatowt (2025) show that model performance can be inflated by artefacts of how inputs are structured rather than by true semantic or causal understanding, and the present finding is a quantitative-modelling analogue of that same warning. Consequently, the SHAP-identified dominance of yoy_scope1_change_pct, dpg_index, and esg_score_0_100 (~77% of ensemble importance) should be read as confirmation that the *model has learned the DPG construction rule*, not as independent evidence that these are the true drivers of real-world greenwashing behaviour.

5.7 Implications for the Proposed Phase 3 Architecture

Taken together, these findings substantiate rather than merely illustrate the core motivation for the proposed DualBranch Contrastive Calibration architecture. The rating-lens dependency finding (5.2) demonstrates empirically why a single-lens "talk" or "walk" proxy is insufficient and why the framework's separation of a Disclosure/Narrative Representation branch from an Operational Reality branch is a necessary design response rather than an arbitrary architectural choice. The unresolved crosswashing reversal (5.3) and the net-zero paradox (5.4) indicate that the underlying construct being modelled is more mechanistically complex than a single binary flag can represent, reinforcing the case for the multi-stage Claim Detection → Operational Evidence Extraction → Dual-Branch Calibration → Independent Verification pipeline proposed in Section 15 of the technical report. Finally, because this dataset contains no visual or report-design data, it cannot speak to the "greenwashing paradox" documented by Jakubczak et al. (2026) — in which commercial multimodal LLMs paradoxically penalise high-performing, visually polished sustainability reporters — but the quantitative DPG baseline established here is precisely the kind of verifiable operational anchor that Branch B of the proposed architecture would need in order to calibrate visual rhetoric against physical reality rather than against aesthetic sophistication alone.

5.8 Caution on Causal Language and Generalisability

None of the associations reported above should be read causally. The dataset is a curated, non-random panel of 30 firms observed repeatedly over 15 years, which violates the independence assumptions of the chi-square and correlation tests applied, and several sub-analyses (e.g., the country-level breakdown, the firm-size direction check) are based on extremely small clusters (in some cases 1–2 firms). The appropriate interpretive register throughout is therefore associative and hypothesis-generating — for example, "elevated ESG scores were associated with rising Scope 1 emissions among flagged firm-years in this sample" — rather than "high

ESG positioning causes emissions increases," and any extension of these findings beyond the Energy, Utilities, and Industrials sectors represented here should be treated as a testable hypothesis for Phase 3, not an established conclusion.

6. Conclusion

6.1 Restatement of the Research Focus

This study examined whether a measurable divergence exists between a corporation's reported ESG narrative ("Talk") and its actual physical emissions trajectory ("Walk"), and whether this Disclosure –Performance Gap (DPG) can function as a reliable statistical indicator for screening corporate sustainability disclosures for greenwashing risk. The analysis synthesised textual, quantitative, and multimodal approaches to AI-driven ESG auditing and tested the resulting claims against a 450 firm-year panel spanning 2010–2024.

6.2 Major Findings

The findings confirm a clear and consistent Talk–Walk divergence: flagged firm-years exhibited markedly higher average ESG scores (64.4 versus 55.6) alongside rising, rather than falling, Scope 1 emissions (+2.83% versus −3.82% for non-flagged firms), a pattern consistent with the disclosure–performance gap identified in the literature (Zeng et al., 2025). The analysis further confirmed that greenwashing is not a fixed corporate attribute but a construct contingent on the measurement lens chosen: the same ESG score, paired with two different performance proxies, produced correlations with the greenwashing flag ranging from $r = 0.103$ to $r = 0.480$ — a substantial divergence that provides direct empirical confirmation of the rating-lens dependency previously raised only conceptually in the literature (Kumar & Sieradzki, 2026). At the same time, the apparently strong predictive performance of the ensemble classifiers (ROC-AUC > 0.99, 97% accuracy) was found to be a methodological artefact rather than a genuine finding, driven by target leakage between the engineered DPG features and the label they were used to predict.

6.3 Overall Interpretation

Taken together, these results support the overall interpretation that a greenwashing flag should be understood strictly as a statistical risk-screening signal, not as a legal finding of fraud, deception, or intentional corporate wrongdoing. A firm's sustainability narrative frequently operates independently of its underlying physical emissions trajectory, and the value of an algorithmic lens lies in flagging where marketing and operational reality diverge, not in adjudicating corporate intent.

6.4 Contribution

The study's principal contribution is the construction of a transparent, quantitative DPG baseline that moves the literature beyond sentence-level claim detection, which can identify whether a statement is environmental but not whether it is misleading. Building on this

baseline, the study proposes a conceptual Split iXBRL-Aware Agentic Verification Framework, in which a high-recall agent extracts machine-readable financial tags and a focused verification agent aligns narrative claims with audited capital expenditures, together with a Dual-Branch Contrastive Calibration model that decouples visual and narrative rhetoric from verifiable operational performance. The research also contributes a policy-relevant insight: mature disclosure infrastructure and external audits appear to function as enforcement-magnifying mechanisms that make inconsistencies more visible, rather than as preventative shields against greenwashing itself, and it demonstrates empirically how sensitive greenwashing measurement is to the choice of underlying metric.

6.5 Implications

These findings carry clear implications for how AI should be positioned within ESG compliance. Given the "volitional vacuum" and "garbage-in, garbage-out" risks inherent in automated disclosure analysis, AI-based tools should function as risk-screening and evidence-prioritisation instruments rather than autonomous compliance auditors, with high-risk cases subject to mandatory algorithmic due diligence, human sign-off, and third-party physical verification consistent with emerging requirements under the CSRD/ESRS regime. The proposed framework is accordingly best understood as a decision-support mechanism intended to direct human audit and compliance resources toward the cases of greatest concern, rather than to substitute for the judgement such determinations require.

6.6 Limitations

The study is nonetheless subject to important limitations. The reported ROC-AUC above 0.99 must not be read as evidence of real-world detection capability, since it reflects circularity in the DPG construction rather than genuine predictive validity; the dataset was confined to firms operating under US and European regulatory regimes; it contained no visual or report-design features, so the proposed visual–operational contrastive calibration could not be empirically tested in this phase; and the practical deployment of an iXBRL-aware multimodal verification framework remains constrained by the current state of corporate reporting infrastructure, a constraint expected to ease as machine-readable disclosure mandates such as the European Single Access Point are progressively implemented from 2026 onward.

6.7 Future Research

These limitations point directly to the next steps for this research. Future work should conduct leakage-controlled and out-of-time validation — including models that exclude the DPG index outright and models trained on earlier years and tested on later ones — before any claim of genuine predictive capability can be substantiated. Subsequent studies should also test the dual-branch prototype against larger, cross-jurisdictional datasets, extending the analysis to emerging-market disclosure frameworks such as India's Business Responsibility and Sustainability Reporting (BRSR) regime and broader Global South contexts, in order to establish whether the rating-lens sensitivity and jurisdictional effects observed here generalise beyond a US–EU setting.

6.8 Concluding Remarks

In sum, the principal contribution of this study is not a finished or autonomous greenwashing-detection system, but a transparent, quantitative risk-screening baseline that empirically substantiates concerns already raised in the literature regarding rating-lens dependency and narrative bias, and that establishes the structural requirements for the multimodal verification framework to be developed and tested in the next phase of this research.

6.9 Policy Recommendations

The findings of this study point to several concrete actions for the regulators, standard-setters, disclosure preparers, and technology developers involved in ESG reporting and verification.

For regulators and standard-setters (e.g., the bodies overseeing CSRD/ESRS and comparable regimes)

Mandatory algorithmic due diligence requirements should be introduced for any automated system used to screen or flag corporate sustainability disclosures, requiring disclosure of the performance proxy underlying any greenwashing or disclosure–performance metric used. This recommendation follows directly from the finding that the same ESG score produced correlations with the greenwashing flag ranging from r = 0.103 to r = 0.480 depending solely on the choice of performance proxy; without mandated proxy disclosure, two regulators or raters applying nominally similar screening methods could reach materially different risk assessments of the same firm. Regulators should also accelerate the rollout of machine-readable disclosure infrastructure, such as the European Single Access Point, since the practical value of automated verification frameworks of the kind proposed here is constrained until financial and emissions data are consistently available in structured, machine-readable form.

For corporate disclosure preparers

Firms should be required, or strongly encouraged, to report Scope 1–3 emissions trends alongside narrative ESG claims in a standardised, machine-readable format, rather than as prose embedded in unstructured sustainability reports. This addresses the study's finding that flagged firms combine elevated ESG scores with rising, not falling, emissions; standardised co-reporting of these two categories of information would make the disclosure– performance gap directly observable to regulators, investors, and auditors, rather than requiring inference from disparate document formats.

For ESG rating agencies and index providers

Rating methodologies should disclose which operational performance proxy (e.g., carbon intensity versus yearon-year emissions trend) underlies a given ESG or greenwashing-risk score, and should ideally report scores under more than one proxy where feasible. This follows directly from the study's confirmation of rating-lens dependency: a single choice of proxy can more than quadruple the apparent association between disclosure and greenwashing risk, meaning that undisclosed methodological choices materially shape market and regulatory outcomes.

For auditors and third-party verifiers
High-risk cases identified by automated screening tools should be routed to mandatory human review and independent verification against audited capital-expenditure and operational data before any greenwashing determination is made public or acted upon. This recommendation follows from the study's finding that the near perfect predictive accuracy reported by unmodified classifiers was an artefact of target leakage rather than genuine detection capability; treating automated outputs as final determinations, rather than as prioritisation signals requiring verification, risks both false accusations and missed cases.

For developers of AI-based ESG compliance tools
Model developers should adopt leakage-controlled and out-of-time validation as a minimum reporting standard before any predictive claim is presented to regulators, clients, or the public, and should avoid deploying disclosure only models as standalone determinants of greenwashing risk. This is a direct response to the study's demonstration that disclosure-based variables alone yield only marginal genuine predictive power once definitional leakage is removed, indicating that text-and-score-only tools are not yet suitable as autonomous compliance instruments.

For institutional investors and users of ESG ratings
Investors relying on third-party ESG or greenwashing-risk scores should request disclosure of the underlying performance proxy and treat single-proxy scores as provisional rather than definitive, particularly for firms operating in jurisdictions such as the United States and the European Union where this study observed materially different flag rates. This recommendation addresses the risk that investment decisions may otherwise be based on a rating artefact rather than a stable underlying reality.
Collectively, these recommendations are oriented toward the same conclusion reached by the study itself: automated tools of the kind examined here are best positioned as risk-screening and evidence-prioritisation instruments that direct human and regulatory attention toward cases of concern, rather than as autonomous arbiters of corporate compliance.

Research Assistance & Sourcing: This study was conducted with the aid of computational literature-search tools and AI assisted for structuring for the literature review and with AI-assisted coding support within the Google Colab and Kaggle environments used for the statistical and predictive analysis. All literature was cross-verified against open-access academic repositories and primary regulatory sources, and all code and analytical outputs were independently reviewed and tested by the research team. Following these assistive processes, the content, analytical synthesis, statistical interpretation, and citations were critically evaluated, validated, and finalized by the authors.

References

[1] Al-Zoubi, T. M. H., Al-Hailat, O., Alomar, A., & Al-Billeh, T. (2026). AI-driven sustainability reporting and corporate greenwashing: Legal accountability and governance challenges in the ESG era. Sustainability, 18(17), 8661 https://www.mdpi.com/2071-1050/18/17/8661

[1] Baghel, P., & Desai, S. (2025). AI-driven ESG disclosures: A new frontier for greenwashing risk and regulation. International Journal of Creative Research Thoughts, 13(4), 112–124. https://ijcrt.org/papers/IJCRT2506363.pdf

[1] Bergougui, B., Sulimany, H. G. H., & Alharbi, A. A. (2026). Combating ESG greenwashing through AI models. Corporate Social Responsibility and Environmental Management, 33(1), 145–162. https://onlinelibrary.wiley.com/doi/10.1002/csr.70689

[1] Birti, M., Maurino, A., & Osborne, F. (2025). Optimizing large language models for ESG activity detection in financial texts. Proceedings of ICAIF '25, 214–222. https://arxiv.org/abs/2502.21112

[1] Braun, N. H., Ong, K., Mao, R., Cambria, E., & Mengaldo, G. (2026). Enhancing language models for robust greenwashing detection. arXiv:2601.21722. https://arxiv.org/abs/2601.21722

[1] Calamai, T., Balalau, O., Le Guenedal, T., & Suchanek, F. (2025). Detecting greenwashing: A natural language processing literature survey. arXiv:2502.07541. https://arxiv.org/abs/2502.07541

[1] Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting (Corporate Sustainability Reporting Directive, CSRD). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2464&from=EN

[1] Florstedt, J., Fahlbusch, J., & Sontheimer, M. (2025). Detecting greenwashing hints in ESG reports. TU Berlin https://www.researchgate.net/publication/374724773_Detecting_Greenwashing_in_the_Environmental_Social_and_Governance_Domains_Using_Natural_Language_Processing

Working Paper Series.

[2] Hassani, B. K., & Bahini, Y. (2024). Crosswashing in sustainable investing. arXiv:2407.00751. https://arxiv.org/abs/2407.00751

[3] Jakubczak, J. K., Chmielewska-Muciek, D., & Iwanicka, K. (2026). Evaluating multimodal AI for greenwashing detection. Sustainability, 18(1), 236. https://www.mdpi.com/2071-1050/18/1/236

[4] Kumar, K., & Sieradzki, M. (2026). Same firms, different verdicts: ESG rating choice and the measurement of greenwashing. Journal of Business Ethics (forthcoming). https://link.springer.com/journal/10551

[5] Kurfalı, M., Zahra, S., Nivre, J., & Messori, G. (2025). Climate-Eval: A comprehensive benchmark for NLP tasks related to climate change. arXiv:2505.18653. https://arxiv.org/abs/2505.18653

[6] Li, Y., Li, J., Zheng, Y., Zheng, G., & Wu, C. (2025). Machine learning detection of manipulative environmental disclosures in corporate reports. Scientific Reports, 15, Article 45011. https://www.nature.com/articles/s41598-025-29621-y

[7] Ong, K., Mao, R., Varshney, D., Cambria, E., & Mengaldo, G. (2025). Towards robust ESG analysis against greenwashing risks. Proceedings of ACL 2025, 3120–3135. https://aclanthology.org/2025.acl-long.723/

[8] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://www.tandfonline.com/doi/abs/10.2753/MIS0742-1222240302

[9] Rosak-Szyrocka, J. (2026). Artificial intelligence in ESG reporting: A Scopus-based bibliometric analysis. https://www.researchgate.net/publication/407535664_Artificial_Intelligence_in_ESG_Reporting_A_Scopus-Based_Bibliometric_Analysis_and_Conceptual_Model_for_Data-Driven_Decision_SupportProduction Engineering Archives, 32(3), 430–441.

[10] Schmoll, J., & Jatowt, A. (2025). Automated analysis of sustainability reports using LLMs. arXiv:2512.24289. https://arxiv.org/abs/2512.24289

[11] Shankar, R., & Qian, X. (2026). Automated detection of greenwashing in Indian corporate sustainability reports using NLP. SSRN 5294480. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5294480

[12] Singh, S., & Singh, M. (2024). Algorithmic criminal liability in greenwashing: Comparing India, United States, and European Union. HPNLU Journal of Law, Business and Economics, 3, 51–68. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5504540

[13] Sulkowski, A. J. (2025). AI, ESG, and law: Potential, limitations, and strategies concerning artificial intelligence in sustainability reporting. Texas Environmental Law Journal (forthcoming). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4972787

[14] Vinella, A., Capetz, M., Pattichis, R., Chance, C., Ghosh, R., & Chang, K.-W. (2024). Leveraging language models to detect greenwashing. arXiv:2311.01469. https://arxiv.org/abs/2311.01469

[15] Yang, X., Ren, Z., & Xie, Y. (2024). The effect of artificial intelligence technology on corporate greenwashing level. Warwick Evans Publishing, 9, 88–102. https://wepub.org/index.php/TEBMR/article/view/3090

[16] Zeng, F., Wang, J., & Zeng, C. (2025). An optimized machine learning framework for predicting and interpreting corporate ESG greenwashing behavior. PLOS ONE,20(3),e0316287. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0316287