

A Multimodal LLM Architecture and Triangulated Index for Detecting Rhetorical Drift in Corporate Sustainability Reports

Authors

Abi Cit, Billy Elly, Timileyin Olawale, Jimmy Grace, Kate Bright

Date; October 13, 2026

Abstract

Corporate sustainability reporting has become a critical channel for stakeholder communication, yet the phenomenon of "rhetorical drift"—systematic shifts in language and imagery that obscure deteriorating environmental performance—remains inadequately addressed by existing detection frameworks. This study addresses this gap by developing a multimodal large language model (LLM) architecture integrated with a triangulated index for detecting rhetorical drift in corporate sustainability reports. Drawing on a corpus of 1,247 sustainability reports from 312 firms across seven sectors (2018–2025), we combined vision-language processing of document layouts and imagery with BERT-based linguistic analysis and verified ESG performance metrics. The architecture achieved 89.4% detection accuracy for regulator-sanctioned firms exhibiting rhetorical drift, outperforming text-only baselines by 23.7 percentage points. The triangulated index demonstrated strong convergent validity with stakeholder credibility assessments ($r = .71$). Our findings establish that rhetorical drift operates through coordinated multimodal mechanisms—lexical inconsistency, visual-narrative misalignment, and omission of negative performance indicators—rather than isolated textual signals. The framework provides practitioners with an auditable tool for sustainability governance and advances theoretical understanding of strategic environmental communication as a multimodal phenomenon.

Keywords: rhetorical drift, multimodal LLM, greenwashing detection, sustainability reporting, triangulated index, vision-language models

1. Introduction

1.1 Background

Corporate sustainability reporting has evolved from a voluntary reputational exercise to a regulatory imperative across major markets. The European Union's Corporate Sustainability Reporting Directive (CSRD) and the International Sustainability Standards Board (ISSB) framework now mandate detailed environmental disclosures for thousands of firms . Yet the proliferation of sustainability reporting has been accompanied by growing stakeholder skepticism regarding the authenticity of corporate environmental claims. The phenomenon commonly termed "greenwashing" encompasses a spectrum of practices, from outright misrepresentation to subtler forms of strategic ambiguity and selective disclosure .

A particularly insidious pattern has emerged in recent years: firms experiencing deteriorating environmental performance frequently engage in what we term **rhetorical drift**—a coordinated evolution of language, imagery, and disclosure structure that creates an impression of progress while actual performance stagnates or declines. Research by Chalmers and Klingler-Vidra demonstrates that lexical inconsistency in sustainability reports surged dramatically after 2012, with a pronounced spike between 2019 and 2021 . This inconsistency is not random; it reflects strategic adaptation to shifting stakeholder expectations and regulatory pressures. Recent evidence from US corporate filings shows firms reduced politically salient ESG terminology by nearly 25% between 2024 and 2025 while preserving disclosure substance—a pattern consistent with "ESG camouflage" .

The detection challenge is fundamentally multimodal. Sustainability reports communicate through prose, data tables, photographic imagery, and document architecture (page layout, section prominence, color coding). A firm might maintain positive textual framing while reducing quantitative disclosure, or deploy aspirational imagery of renewable energy while burying emissions increases in dense footnotes. Existing detection approaches have addressed these modalities in isolation, failing to capture the coordinated nature of rhetorical drift .

1.2 Problem Statement

Despite decades of research on greenwashing, existing measurement models struggle to distinguish intentional deception from poor performance or to account for how stakeholders interpret corporate claims . Three specific limitations characterize the current state of knowledge.

First, **methodological fragmentation**. Linguistic approaches relying on sentiment analysis or keyword detection capture textual signals but ignore the visual and structural dimensions of sustainability reports. A firm may use positive environmental language while deploying imagery that subtly undermines its claims—such as showing a small solar installation adjacent to a sprawling fossil fuel facility. Conversely, visual analysis approaches cannot assess the factual accuracy of textual claims .

Second, **temporal blindness**. Most detection frameworks operate on single reports or aggregate across years without modeling the trajectory of rhetorical change. Rhetorical drift is inherently longitudinal—it manifests as systematic shifts in language and imagery that correlate with performance trajectories. Static analysis misses the strategic adaptation that characterizes sophisticated greenwashing .

Third, **validity gaps**. Existing indices often lack external validation against objective performance benchmarks or stakeholder perceptions. A detection tool that flags firms for greenwashing without corroboration from regulatory sanctions or stakeholder assessments provides limited practical utility for governance .

The absence of an integrated, temporally aware, and externally validated framework for rhetorical drift detection constitutes a significant gap in both academic literature and practical governance infrastructure.

1.3 Objectives of the Study

General objective: To design, implement, and validate a multimodal large language model architecture integrated with a triangulated index for detecting rhetorical drift in corporate sustainability reports.

Specific objectives:

1. To identify the multimodal predictors (linguistic, visual, structural) that most accurately signal rhetorical drift in sustainability disclosures.
2. To design a vision-language LLM architecture that jointly processes textual, visual, and layout information from sustainability reports for drift detection.
3. To develop and validate a triangulated index that integrates linguistic deception signals, verified performance data, and stakeholder credibility assessments.
4. To validate the framework against regulator-sanctioned firms and stakeholder perception data across multiple sectors and geographic regions.

1.4 Research Questions

1. What combination of multimodal features—lexical inconsistency, visual-narrative misalignment, structural omission patterns—most accurately predicts rhetorical drift in corporate sustainability reports?
2. How does the proposed multimodal LLM architecture compare to text-only and single-modal baselines in detecting rhetorical drift?
3. What is the convergent validity of the triangulated index against regulatory sanctions and stakeholder credibility assessments?
4. What implementation considerations (computational requirements, data availability, interpretability) affect the practical deployment of multimodal rhetorical drift detection?

1.5 Significance of the Study

For practitioners and administrators: The framework provides an auditable tool for identifying sustainability reports warranting enhanced scrutiny. Compliance officers, ESG analysts, and investment professionals gain a replicable methodology for detecting coordinated multimodal deception that escapes conventional review processes.

For policymakers and regulators: The study offers empirical evidence on the nature and prevalence of rhetorical drift, informing disclosure standards and enforcement priorities. The triangulated index provides a potential foundation for regulatory screening tools.

For academic literature: The research advances theoretical understanding of greenwashing as a multimodal, temporally dynamic phenomenon. It bridges previously disconnected literatures on linguistic deception detection, visual communication, and document understanding.

For future researchers: The open architecture specification and validated index enable replication, extension to new domains, and refinement as reporting practices evolve.

1.6 Scope and Limitations

Scope: The study analyzes English-language sustainability reports from publicly traded firms across seven sectors (energy, materials, industrials, consumer discretionary, consumer staples, financials, utilities) in North America, Europe, and Asia-Pacific, covering reporting years 2018–2025. The corpus includes standalone sustainability reports, integrated annual reports, and dedicated ESG sections within annual reports.

Exclusions: Small and medium enterprises, private firms, and non-English reports are excluded due to data availability constraints. Financial-sector firms are included only when they report on their own operational environmental performance, not their financed emissions.

Limitations: Several limitations should be acknowledged upfront. First, the definition of "rhetorical drift" relies on observable disclosure patterns; underlying intent cannot be directly

measured. Second, the validation benchmark (regulator-sanctioned firms) captures only detected violations, potentially underrepresenting undetected drift. Third, computational requirements for multimodal LLM inference may limit accessibility for resource-constrained organizations. Fourth, the corpus, while geographically diverse, overrepresents large-cap firms with established reporting practices.

2. Literature Review

2.1 Conceptual Review

Greenwashing and Rhetorical Drift. The term "greenwashing" was coined by environmentalist Jay Westerveld in 1986 to describe hotel practices of promoting towel reuse while generating substantial waste elsewhere. Academic conceptualization has evolved from a focus on outright false claims to encompass selective disclosure, symbolic management, and strategic ambiguity . We define rhetorical drift as the systematic, temporally correlated evolution of language, imagery, and disclosure structure that creates a favorable impression of environmental performance while objective indicators deteriorate or stagnate. This definition distinguishes drift from isolated exaggerations by emphasizing coordinated, directional change across multiple communicative modalities.

Multimodal Document Understanding. Sustainability reports are complex multimodal artifacts. They combine natural language prose, structured data tables, photographic imagery, infographics, and document architecture (page layout, section ordering, visual prominence). Document understanding research has progressed from OCR-based pipelines to vision-language models that jointly process text, layout, and visual elements . LayoutLMv3 and similar architectures demonstrate that 2D positional embeddings and cross-modal attention substantially improve performance on visually rich documents . Recent work on visual prompting and fine-grained grounding enables models to reason over user-specified document regions, enhancing interpretability and verification .

Lexical Consistency and Deception Detection. Linguistic approaches to greenwashing detection leverage natural language processing to identify markers of deceptive communication. Lexical inconsistency—shifts in terminology used to describe the same concept—has been identified as a key mechanism explaining ESG backlash . Stylometric features including readability, concreteness, and sentiment polarity provide additional signals. BERT-based models fine-tuned on sustainability corpora achieve strong performance on classification tasks relevant to deception detection . However, purely textual analysis misses the visual and structural dimensions of sustainability communication.

2.2 Theoretical Framework

This study integrates three theoretical perspectives to explain rhetorical drift and guide detection architecture design.

Prospect Theory. Kahneman and Tversky's prospect theory posits that decision-makers are loss-averse, weighting potential losses more heavily than equivalent gains. Applied to corporate disclosure, prospect theory predicts that firms facing reputational or regulatory losses from poor environmental performance will engage in risk-seeking communication strategies—including rhetorical drift—to avoid acknowledging deterioration. The theory explains why firms intensify impression management precisely when performance declines, and why disclosure patterns may shift more rapidly than underlying performance .

Signaling Theory. Spence's signaling theory addresses information asymmetry between firms and stakeholders. Sustainability reports function as signals of environmental commitment, but signaling is susceptible to mimicry and deception. When the cost of producing misleading signals is low relative to the benefits of favorable stakeholder responses, greenwashing becomes rational. Signaling theory guides our focus on the credibility mechanisms—third-party verification, quantitative specificity, stakeholder corroboration—that distinguish honest from deceptive signals .

Legitimacy Theory. Suchman's legitimacy theory explains organizational behavior as motivated by the desire to maintain social acceptance. Firms whose environmental practices threaten legitimacy may engage in symbolic management, decoupling disclosure from action. Rhetorical drift represents a sophisticated form of legitimacy management that maintains surface-level conformity to evolving stakeholder expectations while minimizing substantive change .

2.3 Empirical Review

The empirical literature on greenwashing detection has advanced significantly but remains characterized by methodological fragmentation.

Linguistic Detection Studies. Chalmers and Klingler-Vidra analyzed over 6,500 sustainability reports from 320 major firms (2000–2021), finding a marked increase in lexical inconsistency beginning around 2012 and accelerating through 2019–2021 . Their NLP approach combined TF-IDF cosine similarity, stylometric analysis, and sentiment metrics. While providing valuable longitudinal evidence, the study relied exclusively on textual signals and did not incorporate visual or structural document features.

Khandelwal et al. developed a Greenwashing Risk Index using Guided LDA and concreteness scoring on SEC 10-K filings from S&P 50 firms . Their analysis revealed that environmental content constituted only 22.4% of ESG discussion, with governance dominating at 46.4%. The concreteness score of 65.61% suggested that most ESG language was factual and measurable, yielding an average GRI of 2.3%—indicating balanced communication. However, this study was

limited to textual analysis of regulatory filings and did not address standalone sustainability reports or visual communication.

Multimodal Approaches. Morio et al. introduced a benchmark dataset of expert-annotated video advertisements from oil and gas companies, designed to evaluate vision-language models for greenwashing detection . Their baseline experiments showed GPT-4.1 achieving 79% F1 on environmental message detection but only 46% F1 on identifying green innovation framing. This study demonstrated the feasibility of multimodal greenwashing detection while revealing substantial room for improvement, particularly for implicit and culturally embedded framing.

A recent Indonesian study integrated rule-based NLP, claim-verifiability assessment, and multimodal coding across six sustainability reports . Results indicated that promissory language dominated textual content, with most coded units classified as having low verifiability. Cover analysis revealed recurrent combinations of nature imagery, community actors, and green color salience—producing strong visual narratives of responsibility. This study's diagnostic framework conceptualized greenwashing as misalignment among linguistic projection, evidential completeness, and multimodal legitimacy. However, the small sample size (six reports) limits generalizability.

Triangulated and Validated Frameworks. Okeke and Rahim introduced the Greenwashing Index (GWI), integrating BERT-based linguistic deception detection, verified performance data (ESG scores, EPA violations), and stakeholder credibility assessments . The GWI identified 83% of regulator-sanctioned firms and correlated strongly with stakeholder-rated greenwashing severity ($r = .71$). Cluster analysis revealed four behavioral archetypes, with NGOs significantly more punitive than investors. This study represents the most comprehensive validation effort to date but focuses on aggregate greenwashing rather than temporal rhetorical drift and does not incorporate multimodal document analysis.

2.4 Research Gap

Despite these advances, no validated framework exists that simultaneously addresses three critical dimensions of rhetorical drift detection. First, **multimodal integration**: existing approaches analyze text or visuals in isolation, missing coordinated cross-modal signals. Second, **temporal dynamics**: detection frameworks typically analyze single reports or aggregate across years without modeling the trajectory of rhetorical change. Third, **triangulated validation**: few indices validate against both objective performance benchmarks and stakeholder perceptions.

The proposed research fills this gap by developing a multimodal LLM architecture that jointly processes textual, visual, and layout information; models rhetorical drift as a temporal phenomenon; and validates the resulting index against regulatory sanctions and stakeholder assessments. This integrated approach enables detection of sophisticated greenwashing that exploits the multimodal and longitudinal nature of modern sustainability reporting.

3. Methodology

3.1 Research Design

This study employs a **design science research** approach combined with **retrospective longitudinal analysis** and **prospective validation**. Design science is appropriate for developing and evaluating artifacts—here, the multimodal LLM architecture and triangulated index—that address practical problems. The retrospective component analyzes historical sustainability reports to identify rhetorical drift patterns and train detection models. The prospective component validates the framework against external benchmarks.

The design proceeds through four phases: (1) corpus construction and annotation, (2) multimodal architecture development, (3) triangulated index construction, and (4) validation against regulatory and stakeholder benchmarks. This sequence enables iterative refinement informed by intermediate results.

3.2 Study Area / Population

The target population comprises publicly traded corporations that publish English-language sustainability reports and have verifiable environmental performance data available through regulatory databases or third-party ESG assessments. The sampling frame draws from firms listed on major exchanges (NYSE, NASDAQ, LSE, Euronext, TSE, ASX) with market capitalization exceeding \$2 billion and at least three consecutive years of sustainability reporting between 2018 and 2025.

3.3 Sample Size and Sampling Technique

Sample size: The final corpus includes 1,247 sustainability reports from 312 firms.

Sampling method: Stratified purposive sampling was employed to maximize variance across:

- **Sector** (7 strata: energy, materials, industrials, consumer discretionary, consumer staples, financials, utilities)
- **Geographic region** (3 strata: North America, Europe, Asia-Pacific)
- **Environmental performance trajectory** (3 strata: improving, stable, deteriorating, based on third-party ESG scores)
- **Regulatory status** (2 strata: sanctioned for environmental violations, non-sanctioned)

Stratification ensures representation of firms across the performance spectrum, including both high-performing firms that might engage in "defensive" rhetorical maintenance and low-performing firms with stronger incentives for drift.

Justification: The sample size of 312 firms provides adequate statistical power for detecting medium effect sizes ($f^2 = 0.15$) in regression analyses with up to 15 predictors at $\alpha = .05$ and power

= .80. The longitudinal structure (approximately 4 reports per firm on average) enables within-firm fixed-effects modeling that controls for stable firm characteristics.

3.4 Data Collection Methods

Data sources:

1. **Sustainability reports:** Downloaded from corporate websites, ESG databases (Corporate Register, GRI Standards Database), and SEC EDGAR for firms filing integrated reports.
2. **Verified performance data:** Environmental performance metrics obtained from EPA Enforcement and Compliance History Online (ECHO), European Pollutant Release and Transfer Register (E-PRTR), and third-party ESG ratings (MSCI, Sustainalytics, Refinitiv).
3. **Regulatory sanctions:** Enforcement actions from EPA, European Environment Agency, and national environmental regulators.
4. **Stakeholder assessments:** Credibility ratings from CDP (formerly Carbon Disclosure Project), Sustainalytics controversy scores, and NGO watchdog reports.

Data types extracted:

- **Textual:** Full report text with section boundaries (environmental, social, governance sections identified via section headers and table of contents parsing).
- **Visual:** Page-level images (300 DPI) with bounding boxes for photographs, infographics, and charts.
- **Layout:** PDF structure metadata including page dimensions, text block coordinates, font information, and reading order.
- **Quantitative:** Emissions data (Scope 1, 2, 3), water usage, waste generation, compliance violations, and ESG scores.

Time periods: Reports published between January 1, 2018 and December 31, 2025.

Simulated data: For a subset of firms ($n = 47$), granular operational data required for performance-drift correlation was not publicly available. For these cases, emissions trajectories were imputed using sector-median trends and firm-specific financial proxies (revenue, capital expenditure), following the imputation methodology validated by Khandelwal et al. . Sensitivity analyses confirmed that results were robust to exclusion of imputed cases.

3.5 Research Instruments

Software and libraries:

- **PyTorch 2.1** and **HuggingFace Transformers 4.35** for model implementation

- **LayoutLMv3** (fine-tuned on sustainability document corpus) for layout-aware text extraction
- **InternVL2-2B** as base vision-language model with MoE connector adaptation following DocRouter architecture
- **BERT-base-uncased** fine-tuned for linguistic deception classification
- **spaCy 3.7** for linguistic preprocessing and dependency parsing
- **OpenCV 4.8** for image feature extraction
- **scikit-learn 1.3** for statistical analysis and baseline models

Preprocessing steps:

1. **PDF parsing:** Document structure extracted using PyMuPDF, preserving page layout, text blocks, and image regions.
2. **OCR correction:** LayoutLMv3 OCR outputs validated against embedded text layers, with manual correction for critical sections.
3. **Section segmentation:** Reports segmented into environmental, social, governance, and introductory sections using a fine-tuned section classifier.
4. **Image classification:** Photographs and infographics classified into categories (nature imagery, industrial facilities, people, abstract graphics) using a ResNet-50 classifier trained on 5,000 manually labeled images.
5. **Tokenization:** Text tokenized using the BERT WordPiece tokenizer for linguistic analysis and the InternVL2 tokenizer for multimodal processing.

The system design follows a **triangulated architecture** that processes each sustainability report through three parallel analytical pathways—linguistic, visual-structural, and performance—whose outputs converge in a fusion layer for drift detection. This design draws on Okeke and Rahim's triangulation principle, extending it from measurement constructs to computational architecture.

3.6 Validity and Reliability

Content validity: The feature set for rhetorical drift detection was validated by a panel of five experts in sustainability reporting, ESG analysis, and corporate communication. Experts rated each feature's relevance to drift detection on a 5-point scale; items with mean ratings below 3.5 were excluded. The final feature set achieved a content validity index (CVI) of 0.89.

Predictive validity: The framework's predictions were validated against regulatory sanctions (binary outcome) and stakeholder credibility assessments (continuous outcome). Area under the ROC curve (AUC) and correlation coefficients served as validity indicators.

Inter-rater reliability: A subset of 150 reports was independently coded by two trained research assistants for visual-narrative alignment and omission patterns. Cohen's κ was 0.81 for visual alignment coding and 0.76 for omission coding, indicating substantial agreement.

Construct reliability: Internal consistency of the triangulated index was assessed using Cronbach's α across the three component dimensions. The α coefficient was 0.84, exceeding the conventional 0.70 threshold.

3.7 Data Analysis Techniques

Multimodal LLM Architecture:

The core detection model integrates three components:

1. **Linguistic pathway:** BERT-based encoder fine-tuned on a labeled corpus of deceptive and non-deceptive sustainability claims. Outputs include deception probability scores for each report section and lexical consistency metrics (TF-IDF cosine similarity across years).
2. **Visual-structural pathway:** InternVL2-2B with modified connector (following DocRouter) processes page images and layout information. Attention-gating mechanisms direct focus to environmentally relevant regions. Outputs include visual-narrative alignment scores and document structure metrics (section prominence, disclosure completeness).
3. **Fusion and drift detection:** Outputs from linguistic and visual pathways are fused with performance trajectory data using a gated cross-attention mechanism. The model produces a drift probability score for each report-year observation.

Triangulated Index:

The Rhetorical Drift Index (RDI) combines three validated components:

- **Linguistic Deception (LD):** BERT-based probability of deceptive linguistic patterns, adjusted for lexical consistency trends.
- **Performance-Disclosure Gap (PDG):** Divergence between disclosed environmental performance and verified external metrics, normalized by sector and firm size.
- **Stakeholder Credibility Deficit (SCD):** Inverse of stakeholder credibility assessments, aggregated across CDP, Sustainalytics controversy scores, and NGO reports.

$$\text{RDI} = 0.40 \times \text{LD} + 0.35 \times \text{PDG} + 0.25 \times \text{SCD}$$

Weights were determined through factor analysis of component correlations with the validation criterion (regulatory sanctions), following the methodology established by Okeke and Rahim .

Performance metrics:

- **Accuracy, precision, recall, F1** for binary classification (sanctioned vs. non-sanctioned firms)
- **AUC-ROC** for threshold-independent discrimination
- **Mean average precision (MAP)** for ranking firms by drift severity
- **Feature importance** via SHAP values and permutation importance

Cross-validation: Leave-one-firm-out cross-validation (312 folds) ensures that model performance is assessed on entirely held-out firms, preventing information leakage across a firm's reports. Temporal validation uses reports from 2024–2025 as a holdout test set, with models trained on 2018–2023 data.

Statistical comparison: Model performance compared against baselines using paired t-tests with Bonferroni correction for multiple comparisons. Effect sizes reported as Cohen's d.

3.8 Ethical Considerations

This study analyzes publicly available corporate documents and third-party data sources. No human subjects were directly involved, and no protected health information (PHI) or personally identifiable information (PII) was accessed. The research protocol was submitted to the Institutional Review Board (IRB) of the University of Jos, which determined that the study qualified for exemption under 45 CFR 46.104(d)(4) as research involving only publicly available data. All data handling followed institutional data governance policies. Findings are reported at the firm level, which is standard practice in ESG research given that sustainability reports are public documents intended for stakeholder scrutiny. No individual employees or private individuals are identified.

4. Results

4.1 Data Presentation

Table 1. Corpus Characteristics by Environmental Performance Trajectory

Indicator	Improving (n=98)	Stable (n=121)	Deteriorating (n=93)
Reports per firm (mean, SD)	4.1 (1.2)	3.9 (1.1)	4.2 (1.3)

Indicator	Improving (n=98)	Stable (n=121)	Deteriorating (n=93)
Report length (pages, mean)	87.3	72.6	94.8
Environmental section length (% of total)	24.7%	21.3%	27.9%
Quantitative disclosure density	0.42 (0.11)	0.38 (0.09)	0.29 (0.13)
Lexical consistency (within-firm, 1-year lag)	0.78 (0.09)	0.74 (0.11)	0.61 (0.14)
Visual-narrative alignment score	0.82 (0.08)	0.79 (0.10)	0.64 (0.12)

Note: Quantitative disclosure density = proportion of environmental section comprising quantified metrics. Visual-narrative alignment = correlation between image valence and performance trajectory.

Table 1 presents corpus characteristics stratified by environmental performance trajectory. Deteriorating firms published longer environmental sections (27.9% of total report vs. 21.3% for stable firms) yet exhibited lower quantitative disclosure density (0.29 vs. 0.38). Lexical consistency was markedly lower for deteriorating firms (0.61) compared to improving firms (0.78), indicating greater terminology instability. Visual-narrative alignment scores followed a similar pattern, with deteriorating firms showing weaker alignment between image valence and actual performance.

Table 2. Rhetorical Drift Features by Regulatory Status

Feature	Sanctioned (n=67)	Non-Sanctioned (n=245)	t- statistic	p- value
LD probability	0.73 (0.14)	0.41 (0.17)	8.42	<.001
PDG score	0.68 (0.19)	0.34 (0.16)	7.91	<.001

Feature	Sanctioned (n=67)	Non-Sanctioned (n=245)	t- statistic	p- value
SCD score	0.71 (0.21)	0.38 (0.18)	7.23	<.001
RDI (composite)	0.71 (0.12)	0.37 (0.14)	9.15	<.001

Note: LD = Linguistic Deception; PDG = Performance-Disclosure Gap; SCD = Stakeholder Credibility Deficit.

Table 2 presents feature comparisons between regulator-sanctioned and non-sanctioned firms. All three component dimensions of the RDI showed statistically significant differences, with sanctioned firms exhibiting substantially higher scores. The composite RDI demonstrated the strongest discrimination ($t = 9.15$, $p < .001$), supporting the triangulated approach.

4.2 Analysis of Results

Model Performance.

The multimodal LLM architecture achieved **89.4% accuracy** in detecting rhetorical drift among regulator-sanctioned firms (AUC = 0.93, F1 = 0.87, precision = 0.85, recall = 0.89). This substantially outperformed all baselines:

- **Text-only BERT:** 65.7% accuracy ($\Delta = +23.7$ percentage points, Cohen's $d = 1.42$)
- **Visual-only CNN:** 58.2% accuracy ($\Delta = +31.2$ percentage points, $d = 1.87$)
- **TF-IDF + logistic regression:** 61.3% accuracy ($\Delta = +28.1$ percentage points, $d = 1.63$)
- **Static multimodal (no temporal):** 76.8% accuracy ($\Delta = +12.6$ percentage points, $d = 0.87$)

Paired t-tests confirmed that the multimodal architecture significantly outperformed all baselines at $p < .001$ with Bonferroni correction.

Feature Importance.

SHAP analysis identified the top predictors of rhetorical drift:

1. **Lexical inconsistency across years** (mean $|\text{SHAP}| = 0.142$): Firms engaged in drift systematically shifted terminology, particularly reducing use of specific regulatory terms (e.g., "carbon neutral," "net zero") in favor of vague aspirational language.

2. **Visual-narrative misalignment** (mean $|\text{SHAP}| = 0.118$): Drift was characterized by increased deployment of positive nature imagery despite deteriorating quantitative indicators.
3. **Quantitative disclosure reduction** (mean $|\text{SHAP}| = 0.097$): Drifting firms reduced granular emissions data while increasing narrative length.
4. **Section restructuring** (mean $|\text{SHAP}| = 0.073$): Environmental content was frequently moved from prominent early sections to appendices or footnotes.
5. **Temporal hedging** (mean $|\text{SHAP}| = 0.068$): Increased use of future-oriented language ("will achieve," "committed to") relative to present-tense performance statements.

Temporal Patterns.

Analysis of drift trajectories revealed that rhetorical drift typically precedes regulatory sanctions by **1.7 years** ($SD = 0.8$), suggesting that multimodal detection can provide early warning. Drift intensified following negative performance events (emissions increases, violations) and persisted for an average of 2.3 reporting cycles before either remediation or sanction.

5. Discussion

5.1 Interpretation

The results demonstrate that rhetorical drift in sustainability reports is a **multimodal, temporally dynamic phenomenon** that cannot be adequately captured by single-modal approaches. The 23.7 percentage point improvement over text-only baselines confirms that visual and structural features carry substantial independent signal for drift detection.

Addressing RQ1 (predictive features): Lexical inconsistency emerged as the strongest single predictor, consistent with Chalmers and Klingler-Vidra's finding that terminology shifts signal credibility problems. However, the multimodal architecture revealed that lexical signals operate in coordination with visual and structural changes. Drifting firms did not simply change words; they simultaneously increased aspirational imagery, reduced quantitative specificity, and restructured document architecture to de-emphasize environmental performance. This coordinated pattern explains why text-only approaches, however sophisticated, achieve only moderate accuracy.

Addressing RQ2 (architecture comparison): The fusion of linguistic and visual pathways via gated cross-attention proved essential. Ablation experiments showed that removing the visual pathway reduced accuracy by 18.2 percentage points, while removing temporal modeling reduced accuracy by 12.6 percentage points. The attention-gating mechanism from DocRouter enabled the

model to focus on environmentally relevant document regions, avoiding distraction from unrelated visual content.

Addressing RQ3 (validation): The triangulated index demonstrated strong convergent validity with both regulatory sanctions ($AUC = 0.93$) and stakeholder credibility assessments ($r = .71$), aligning with the validation results reported by Okeke and Rahim . The three components contributed unique variance: linguistic deception captured intent signals, performance-disclosure gap captured action signals, and stakeholder credibility deficit captured perception signals. This triangulation addresses the persistent challenge of distinguishing intentional deception from poor performance or communication incompetence.

Theoretical extension: The findings extend legitimacy theory by demonstrating that symbolic management in sustainability reporting operates through coordinated multimodal mechanisms. Firms do not simply "talk differently" when facing legitimacy threats; they deploy a comprehensive communicative strategy spanning text, image, and document structure. Prospect theory's prediction of loss-averse, risk-seeking communication behavior is supported by the intensification of drift following negative performance events.

5.2 Implications

Academic implications: This study introduces rhetorical drift as a distinct construct from static greenwashing, characterized by temporal directionality and multimodal coordination. The finding that drift precedes sanctions by 1.7 years establishes a temporal dimension to greenwashing detection previously unrecognized. The multimodal LLM architecture provides a methodological template for analyzing strategic communication as an integrated phenomenon rather than a collection of independent signals.

Practical implications for administrators and analysts:

1. **Monitoring metrics:** Organizations should track lexical consistency (year-over-year terminology stability), visual-narrative alignment (correlation between image valence and performance data), and quantitative disclosure density as early warning indicators.
2. **Detection cadence:** Given the 1.7-year lead time, annual screening using the RDI framework can provide actionable intelligence before regulatory action.
3. **Threshold guidance:** Firms with RDI scores exceeding 0.55 (75th percentile of non-sanctioned firms) warrant enhanced scrutiny. Scores above 0.65 indicate high drift probability.
4. **Interpretability requirements:** The attention-gating and provenance-tracking features of the architecture enable auditors to trace detection decisions to specific document regions, satisfying regulatory transparency requirements.

For policymakers: The findings support disclosure standards that require quantitative environmental metrics alongside narrative discussion and penalize systematic terminology instability. The EU AI Act's transparency obligations for high-risk systems provide a regulatory template for governing automated greenwashing detection tools.

5.3 Limitations

1. **Sample and generalizability:** Despite stratification, the corpus overrepresents large-cap firms with established reporting practices. Findings may not generalize to SMEs or firms in emerging markets with different reporting cultures.
2. **Validation criterion limitations:** Regulatory sanctions capture only detected violations. Firms engaging in undetected drift may be misclassified as non-drifters, attenuating observed effect sizes.
3. **Imputed performance data:** For 47 firms (15.1% of sample), granular performance data required imputation, introducing measurement error that may affect PDG scores for those cases.
4. **Temporal stability assumption:** The model assumes that relationships between multimodal features and drift remain stable over time. As reporting practices evolve, feature importance may shift.
5. **Intent inference:** The framework detects patterns consistent with strategic deception but cannot establish intent. Some detected drift may reflect organizational learning or communication strategy evolution rather than deliberate misrepresentation.

5.4 Future Research Directions

1. **Extension to different reporting regimes:** Apply the framework to CSRD-mandated reports, integrated reports following ISSB standards, and voluntary reports to assess how regulatory context shapes drift patterns.
2. **Real-time monitoring architecture:** Develop streaming detection capabilities that update drift assessments as new reports are published, enabling continuous governance oversight.
3. **Causal analysis of drift triggers:** Use event study methodology to identify specific performance events, regulatory changes, or stakeholder actions that precipitate rhetorical drift.
4. **Cross-cultural generalizability:** Examine whether multimodal drift signals vary across linguistic and cultural contexts, particularly in Asian and Latin American reporting environments where visual communication norms differ.
5. **Human-AI collaboration:** Study how auditors and analysts interact with drift detection outputs, optimizing interface design for interpretability and trust calibration.

6. Conclusion

This study developed and validated a multimodal large language model architecture integrated with a triangulated index for detecting rhetorical drift in corporate sustainability reports, achieving 89.4% detection accuracy for regulator-sanctioned firms and demonstrating a 1.7-year average early warning capability. The framework establishes that rhetorical drift is fundamentally multimodal—operating through coordinated changes in language, imagery, and document structure—and temporally dynamic, intensifying in response to performance deterioration. By integrating linguistic deception detection, verified performance data, and stakeholder credibility assessments, the triangulated index transforms greenwashing detection from subjective suspicion into auditable diagnosis. For administrators and governance professionals, the framework offers a replicable methodology for identifying reports warranting enhanced scrutiny, with specific metrics (lexical consistency, visual-narrative alignment, disclosure density) and thresholds for practical deployment. As sustainability reporting continues its evolution from voluntary communication to regulated disclosure, tools that can detect strategic manipulation across modalities and time will become essential infrastructure for maintaining the integrity of environmental governance.

References

1. Okeke, A. (2026). Measuring greenwashing: Validating constructs for corporate environmental misrepresentation. *Innovation and Green Development*, 5(3), 100363.
2. Okeke, A., & Rahim, L. J. (2026). Deconstructing deception: A triangulated greenwashing index for precision sustainability governance. *Organization & Environment*. <https://doi.org/10.1177/10860266261477310>
3. Morio, G., Rowlands, H., Stambach, D., Manning, C. D., & Henderson, P. (2025). A multimodal benchmark for framing of oil & gas advertising and potential greenwashing detection. *Advances in Neural Information Processing Systems*, 38.
4. Chalmers, A. W., & Klingler-Vidra, R. (2023, November 22). Inconsistent language can undermine the credibility of corporate social responsibility reports. *LSE Business Review*.
5. Artiga Gonzalez, T., Calluzzo, P., & Gopal, B. (2025). *Sentiment and sustainability: How the language of U.S. corporate filings reveals divergent paths for management priorities on DEI and environment* (ISF White Paper). The Institute for Sustainable Finance.
6. TRR 266 Accounting for Transparency. (2026). *Staying under the radar: The case of sustainability disclosures in the US* (No. 244).
7. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F., & others. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*.
8. Khandelwal, V., Mahadeshwar, S., & Botsa, P. P. (2026). *Decoding ESG: Measuring clarity and greenwashing in corporate disclosures* [Project repository]. Frankfurt School of Finance and Management.
9. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2025). DocRouter: Prompt guided vision transformer and mixture of experts connector for document understanding. *Information Fusion*, 122, 103206.
10. VIP-Doc Team. (2026). VIP-Doc: Visual prompts guide fine-grained document understanding for reader friendly VLLM. *Expert Systems with Applications*, 128456.
11. Garuda Research Team. (2026). Multimodal analysis of corporate sustainability reports: A comparative corpus-assisted approach. *Garuda Digital Repository*.
12. Garuda Research Team. (2026). Effectiveness of AI-based audit instruments in detecting greenwashing: A systematic literature review. *Garuda Digital Repository*.

13. Millani. (2024). *8th annual ESG disclosure study*. TMX Group.
14. Sage Journals. (2026). Deconstructing deception: A triangulated greenwashing index for precision sustainability governance. *Organization & Environment*.
15. University of Cumbria Research Portal. (2026). Deconstructing deception: A triangulated greenwashing index for precision sustainability governance.
16. Wiley Online Library. (2026). The language of greenwashing: SDG omission and opportunity-oriented environmental tone as alert metrics in green bond disclosures. *Corporate Social Responsibility and Environmental Management*.
17. Faisal, M., Haryanto, A., & Nugroho, A. (2025). Evaluating multimodal AI for greenwashing detection: A comparative analysis of ChatGPT, Claude, and Gemini in ESG reports. *Sustainability*, 18(1), 236.