Greenwashing in Biodiversity Reporting: The Role of Financial Risk Exposure and Conservation Policy Stringency

Felipe Suárez–Giri[a], Silvia Mayoral Blaya[b], Rosa Rodríguez Lopez[b]

[a] *Universidad Politécnica de Madrid, Center for Innovation and Technology for Development, C. de José Gutiérrez Abascal, 2, 28006 Madrid, Spain*
[b] *Universidad Carlos III de Madrid, Department of Business Administration, C/ Madrid 126, 28903 Getafe, Madrid, Spain*

Abstract

The accelerating biodiversity crisis poses systemic risks to financial markets, heightening the need for accurate assessments of firms' biodiversity-related financial risks (BRFR). However, the complexity of ecosystem dynamics and the absence of standardized metrics create fertile conditions for greenwashing, which can systematically distort risk evaluation and pricing. This paper leverages state-of-the-art Natural Language Processing techniques to quantify biodiversity-related greenwashing in sustainability reports and examines its relationship to different sources of institutional pressure. Focusing on Fortune Global 500 firms over the period 2017–2023, we analyze two distinct forms of greenwashing: "cheap talk" (reliance on vague commitments) and "positive talk" (overemphasis on optimistic disclosures). We find that cheap talk is positively associated with firms' exposure to BRFR. Conversely, positive talk is positively associated with the strength of biodiversity protection policies in firms' home countries. Both practices intensified around the Kunming-Montreal conservation policy milestones (2021-2022), with cheap talk rising first alongside growing investor scrutiny, and positive talk strengthening later, as biodiversity expectations became more broadly institutionalized. Our findings carry significant implications for financial actors, as greenwashing can distort the assessment of BRFR, potentially lead-

Email addresses: felipe.suarez@upm.es (Felipe Suárez–Giri), smayoral@emp.uc3m.es (Silvia Mayoral Blaya), rrlopez@emp.uc3m.es (Rosa Rodríguez Lopez)

ing to capital misallocation and sudden repricing shocks. This highlights the need for more rigorous reporting standards in this field.

Keywords: Biodiversity, Greenwashing, ESG, NLP, Sustainable Finance

1. Introduction

Biodiversity loss has emerged as a critical source of systemic global risk, with scientists warning that we are on a trajectory toward our planet's sixth mass extinction—the first driven by human activity—[11, 32]. The financial materiality of biodiversity loss has gained substantial recognition in recent years following key institutional milestones, including the UN's adoption of the Kunming Declaration in 2021 and the Global Biodiversity Framework in 2022. This heightened awareness is transforming financial markets, as asset managers, institutional investors, and central banks increasingly view biodiversity loss as a critical source of systemic risk comparable to climate change [17, 2, 38, 19]. The recent development of various biodiversity reporting standards and frameworks—most notably the establishment of the Task Force on Nature-Related Financial Disclosures (TNFD)—has further institutionalized biodiversity considerations as financially material factors requiring formal disclosure [15].

Biodiversity loss generates multiple, interlinked forms of financial risk. Physical risks emerge when ecosystem degradation disrupts critical services on which firms rely—e.g. pollination, water regulation, soil formation, fish stock regeneration, disease regulation, and coastal protection—leading to supply chain failures, input cost inflation, and operational disruptions [12, 18]. Transition risks emerge as governments tighten biodiversity protection regulations, markets shift toward sustainable products, and investors demand higher standards of environmental performance (ibid.). Meanwhile, reputational risks intensify as civil society demands greater corporate accountability for biodiversity impacts. Together, these risks can erode firm value, increase capital costs, and transmit shocks across sectors and borders, creating systemic vulnerabilities in financial markets [38].

The effective pricing of biodiversity-related financial risks (BRFR) depends critically on information quality. Yet, biodiversity poses distinctive challenges for transparent and credible corporate reporting. The inherent complexity of ecosystem dynamics, the absence of standardized metrics, and the relative novelty of biodiversity risks in financial contexts create an en-

vironment particularly conducive to greenwashing [13, 29]. Through the release of vague commitments or overly optimistic disclosures, firms may capture legitimacy benefits while avoiding the significant costs associated with substantive risk mitigation actions [30, 44]. Such practices distort financial markets by obscuring genuine risk exposures, potentially leading to capital misallocation and sudden corrections when risks materialize [36, 38].

Despite the scale and urgency of these challenges, research at the intersection of finance and biodiversity remains strikingly limited. While influential works have emphasized the need to integrate biodiversity risks into financial research [26, 38], recent literature reviews highlight that finance journals have largely overlooked the topic [10, 20, 24], and empirical evidence is still remarkably scarce [25]. Notable exceptions include Garel et al. [17], who introduce a corporate biodiversity footprint measure and document the emergence of a biodiversity risk premium following the Kunming Declaration and the launch of the TNFD, Giglio et al. [18], who develop biodiversity physical and transition risk metrics and demonstrate that equity markets already started to price these risks, and Gjerde et al. [19], who, drawing on a global survey of firms, document the growing recognition of biodiversity as a material financial risk. Yet, virtually no studies have examined greenwashing in the context of biodiversity reporting. De Grefte and De Bruin [13] explicitly underscores this gap, warning about the specific risks associated with biodiversity-related greenwashing and calling for systematic empirical investigation.

This paper addresses this research gap. We leverage state-of-the-art Natural Language Processing (NLP) techniques to measure two distinct types of greenwashing in the sustainability reports of Fortune Global 500 firms, and examine how these practices relate to different sources of institutional pressures. Our findings reveal that cheap talk -the disclosure of vague commitments- is positively associated with firms' exposure to biodiversity-related financial risks (BRFR), while positive talk -the overemphasis on optimistic biodiversity-related disclosures- is positively associated with the conservation policy stringency of firms' headquarters countries. Both forms of greenwashing have grown significantly around the Kunming–Montreal period, suggesting a strategic response to recent institutional milestones in biodiversity protection. These results highlight biodiversity greenwashing as an emerging factor in risk mispricing, reinforcing the urgency for more robust disclosure standards. To our knowledge, this represents the first systematic application of NLP methods to quantify biodiversity-related greenwashing, contributing both to the emerging literature on nature-related financial risks and to the broader

understanding of strategic disclosure practices in corporate sustainability reporting.

The remainder of this paper proceeds as follows: Section 2 introduces the theoretical background and develops the hypotheses; Section 3 describes our NLP approach; Section 4 presents the data, variables, and model specification; Section 5 presents the results; Section 6 addresses robustness checks; and Section 7 concludes.

2. Legitimacy and Signaling Perspectives on Biodiversity Greenwashing

We integrate legitimacy theory with signaling theory to understand greenwashing behavior in biodiversity reporting. Legitimacy theory posits that firms maintain their social license to operate by conforming to stakeholder expectations and societal norms [37]. In financial markets, environmental legitimacy concerns directly impact firm value through multiple channels, including investor confidence, cost of capital, and regulatory scrutiny [14, 43]. When institutional expectations evolve, companies respond by engaging in legitimacy management strategies that can be either substantive (altering actual practices) or symbolic (managing perceptions without altering operations) [3].

Signaling theory complements this framework by explaining how informed parties communicate their qualities or intentions to uninformed parties through observable actions or disclosures. For signals to be effective, they must be costly to imitate [9], and receivers must possess the ability to correctly interpret them [4]. In contexts where assessing signal quality is difficult, firms have incentives to send cheap symbolic signals rather than costly substantive ones [40, 8].

Biodiversity-related financial risks present acute challenges for effective signaling due to three main factors. First, low observability: firms' biodiversity impacts and dependencies operate through multiple channels that are typically highly intricate, location-specific, and often materialize only in the long term [15, 27]. Second, low comparability: unlike carbon emissions, there are no standardized, widely adopted metrics to quantify a firm's biodiversity footprint, limiting benchmarking and cross-firm assessment [18, 26]. Third, low interpretability: the technical complexity of ecosystem dynamics and the novelty of the consideration of BRFR as material mean that most stakeholders still lack the expertise to accurately assess the relevance and quality of

biodiversity-related disclosures [7, 26].

As awareness of BRFR grows, firms face mounting pressure to demonstrate their engagement with the issue to maintain legitimacy [30]. However, the barriers to effective signaling outlined above provide fertile ground for greenwashing, as firms can potentially obtain legitimacy benefits through symbolic disclosures while avoiding the costs associated with substantive risk mitigation actions [13]. In this paper, we analyze how different sources of institutional pressure (financial risk exposure, policy stringency, and international biodiversity agreements) relate to different forms of greenwashing behavior (cheap talk and positive talk).

Based on the theoretical framework outlined above, we develop the hypotheses for this study.

Firms with high exposure to BRFR face growing scrutiny from investors, who increasingly expect assurance on risk management practices [18]. Yet, a critical institutional gap persists: there is little clarity in financial markets on what constitutes effective biodiversity action [23]. Unlike climate risks, where performance expectations are well-established and metrics standardized, biodiversity accountability and measurement frameworks remain fragmented, largely voluntary, and at a premature stage (ibid).

This uncertainty, coupled with the substantial costs of implementing effective risk mitigation measures, creates perverse incentives for firms to engage in cheap talk as a risk-assurance strategy [13]. This practice involves making vague commitments that create an impression of action without specifying concrete plans or measurable targets. General statements highlighting a firm's commitment to take actions towards "minimizing ecosystem impacts" or "protecting biodiversity" can temporarily satisfy stakeholder expectations, while allowing firms maximum flexibility in their actual practices. This strategy is particularly attractive for firms with high exposure to BRFR, as these face higher assurance expectations, but also higher costs of substantive action. In such contexts, the disclosure of vague commitments can allow firms to manage legitimacy pressures without committing to costly operational changes.

H1: Cheap talk is positively associated with exposure to biodiversity-related financial risks

Firms headquartered in countries with stringent biodiversity protection policies face heightened environmental stewardship expectations. These in-

clude stricter rules on the sustainability and traceability of global supply chains, requirements for net biodiversity gains in development projects, and penalties for ecosystem disturbance, among others [26]. Such policies not only mandate compliance, but also reflect prevailing societal values, creating powerful legitimacy pressures from a broad stakeholder base. Even firms with limited exposure to biodiversity risks are affected, as failure to demonstrate alignment can lead to significant reputational costs.

In this context, firms may resort to positive talk as a deliberate sentiment-management strategy. This approach involves selectively emphasizing favorable biodiversity information—such as beneficial ecosystem impacts, emerging business opportunities, or statements using optimistic language—while downplaying content addressing material risks or adverse environmental effects. By shifting the balance toward positive content, firms can project environmental responsibility and signal alignment with policy and societal expectations.

Positive talk differs from cheap talk in both scope and intent. While cheap talk only targets the specificity of commitments made in sustainability reports, positive talk concerns the overall sentiment conveyed across all biodiversity-related mentions. This broader communicative strategy serves a dual purpose. First, it reinforces legitimacy with a wide range of stakeholders, including regulators, NGOs, media, and local communities. Second, it influences investor sentiment by signaling that stricter policies need not undermine profitability or heighten reputational and regulatory risks, but can instead generate opportunities.

H2: Positive talk is positively associated with the stringency of national biodiversity conservation policies

Biodiversity-related financial risks gained unprecedented prominence in financial markets in the period surrounding two key institutional milestones: the UN's adoption of the Kunming Declaration in 2021 and the Kunming-Montreal Global Biodiversity Framework in 2022. In the years leading up to these developments, biodiversity shifted from a niche environmental concern to a material risk consideration. This evolving institutional landscape created new legitimacy pressures for firms across industries, amplifying both investor and societal demands for disclosure [15]. Given the aforementioned challenges for effective signaling in biodiversity-related reporting, we hypothesize that firms responded to these heightened institutional pressures by increasing their

reliance on both cheap talk and positive talk.

H3: Cheap talk and positive talk intensified around the Kunming-Montreal conservation policy milestones (2021–2022).

3. Natural Language Processing Approach

To construct our biodiversity greenwashing metrics, we developed a multi-stage NLP pipeline specifically adapted to the heterogeneous structure of corporate sustainability reports. Our procedure combined automated text parsing and extraction with the fine-tuning of state-of-the-art NLP models on domain-specific data, as illustrated in Figure 1.

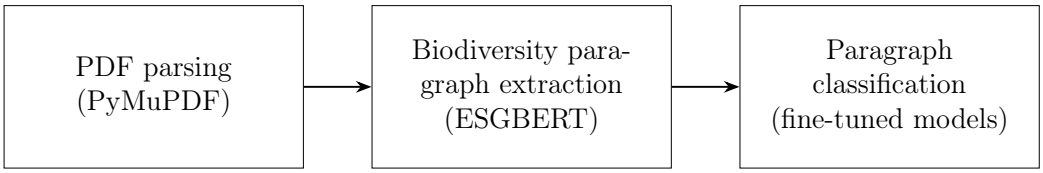


Figure 1: Overview of the text parsing and classification pipeline

First, all sustainability reports were downloaded from firms' websites in PDF format and processed using the Python library PyMuPDF-Fitz. Each document was parsed page by page to retrieve text blocks, with headers and footers excluded using vertical position thresholds. Text blocks were grouped into columns based on horizontal proximity, and paragraph boundaries were reconstructed using vertical spacing rules. To preserve semantic continuity, we implemented a carry-over mechanism for paragraph fragments that spanned across blocks. When the last paragraph on a block did not end in a period, and the first paragraph in the following block began with a lowercase letter —indicating a continuation rather than a new paragraph—, the two were merged. Extensive manual validation confirmed that this strategy effectively minimized fragmentation and preserved semantic continuity, even in reports with highly heterogeneous formatting.

Following text extraction, biodiversity-related content was identified using a version of ESGBERT specifically trained for biodiversity sentence classification [34]. Although the authors of this model report strong performance, we conducted an independent validation to ensure it performs equally well on our sample of Fortune Global 500 reports. We manually classified 1,000

stratified, out-of-sample sentences and compared human labels with model predictions. The classifier achieved similarly high performance on this evaluation set (F1 = 0.945, Accuracy = 0.983, Precision = 0.967, Recall = 0.925), providing confidence that biodiversity sentences are reliably identified in our dataset (see Appendix A for full details).

The ESGBERT-biodiversity model generates a confidence score for each sentence it classifies as biodiversity-related. These scores are highly concentrated near the upper bound (mean = 0.98, median = 1). We conducted a detailed manual review of sentences in the lower tail of the score distribution and found that model performance deteriorates noticeably when confidence scores are low. Based on this assessment, we retain only sentences with scores ≥ 0.9, which account for 93.5% of all extracted sentences. This conservative threshold further ensures that our key variables are not distorted by misclassified text. Paragraphs containing at least one high-confidence biodiversity sentence were flagged as biodiversity-related and used in downstream analysis. All subsequent classification tasks were performed at the paragraph level to capture the broader contextual information necessary for nuanced assessment.

Biodiversity-related paragraphs were classified along three key dimensions: commitment, specificity, and sentiment. For each classification task, we fine-tuned three transformer models based on the DistilRoBERTa architecture. This is a distilled variant of the RoBERTa model developed by Meta AI, which retains most of the original model's performance while substantially reducing computational demands [33].

DistilRoBERTa-base served as our baseline model. We complemented it with two domain-adapted variants, which were built on the same architecture but had been further trained on sustainability and ESG-related corpora. This transfer learning approach has been shown to improve downstream task accuracy in specialized domains [21]. The first variant, ClimateBERT, was trained on a corpus of approximately two million paragraphs from diverse sources, including thousands of corporate sustainability reports [41]. The second, EnvironmentalBERT, was trained on over two million ESG-related sentences, providing a broader environmental and governance context [35].

To fine-tune these models on biodiversity-specific data, we constructed a labeled corpus of 2,000 paragraphs sampled from 300 reports issued by a sample of 70 companies across diverse industries and regions between 2014 and 2024. Each paragraph was manually annotated for commitment, specificity, and sentiment, following the guidelines developed by Bingler et al. [5].

Annotation protocols and illustrative examples are provided in the Appendix B. The label distribution of this fine-tuning dataset is presented in Table 1.

Table 1 - Label Distribution

Task	**Label**	**Label Distribution**
Commitment	Yes/No	1337 / 663
Specific	Yes/No	646 / 691
Sentiment	Pos./Neu./Neg.	1389 / 408 / 203

All models were fine-tuned using a stratified five-fold cross-validation procedure. Model performance was assessed through five-fold averages of weighted precision, recall, F1-score, and unweighted AUROC, consistent with established practices [5]. Results are summarized in Table 2. Per-class performance metrics and out-of-sample pooled confusion matrices are reported in Appendix C. To address class imbalance, we tested focal loss, class weighting, and back-translation data augmentation, but these strategies did not yield meaningful performance gains and were therefore excluded from the final implementation.

ClimateBERT consistently outperformed the other models across all classification tasks. As presented in Table 2, this model achieved the highest scores in all performance metrics considered. Its weighted precision ranged from 89% to 93% across tasks. This model demonstrated a significant improvement over DistilRoBERTa-base and also marginally outperformed EnvironmentalBERT. This performance advantage can likely be attributed to its prior training on large volumes of sustainability-related text.

Hyperparameter optimization was conducted only for ClimateBERT, given its superior performance, while the two underperforming models were not further optimized due to the significant computational and energy costs involved. However, the post-fine-tuning optimization process yielded only marginal performance gains, suggesting that the initial fine-tuning had already achieved near-optimal performance.

Table 2 - Performance Comparison Across Models

Commitment Classification				
Model	Precision	Recall	F1	AUROC
ClimateBERT	0.930	0.929	0.928	0.976
EnvironmentalBERT	0.920	0.920	0.920	0.969
DistilRoBERTa	0.908	0.909	0.908	0.962
Specificity Classification				
Model	Precision	Recall	F1	AUROC
ClimateBERT	0.857	0.856	0.856	0.922
EnvironmentalBERT	0.843	0.841	0.840	0.920
DistilRoBERTa	0.835	0.833	0.833	0.913
Sentiment Classification				
Model	Precision	Recall	F1	AUROC
ClimateBERT	0.902	0.903	0.901	0.968
EnvironmentalBERT	0.900	0.901	0.900	0.967
DistilRoBERTa	0.892	0.894	0.892	0.962

We further assessed classifier reliability through a blinded human audit of model predictions. This external validation provides further support for the robustness of the classifiers used to construct our indicators. We also conducted a qualitative review of all misclassified cases, showing that errors are overwhelmingly concentrated in genuine boundary cases and do not exhibit systematic or directional biases that would mechanically influence our metrics. A detailed account of these validation exercises is presented in Appendix D.

We also evaluated probability calibration for all classifiers using reliability diagrams and Brier scores, and examined performance stability across industries and regions. Across all tasks, classifier behavior is stable across strata, with no evidence of systematic sectoral or geographic degradation. Calibration is strong for the biodiversity, commitment, and sentiment classifiers. The specificity classifier exhibits modest overconfidence, which, given the construction of the Cheap Talk Index -defined in the following section-, induces a conservative bias. This deviation is confined to intermediate predicted probabilities associated with borderline cases and contains relatively

little probability mass, while calibration closely tracks the diagonal at high confidence levels where most classifications occur. Full calibration diagnostics and per-stratum performance results are reported in Appendix E.

4. Case Study

4.1. Data and variables

Our sample comprises firms from the Fortune Global 500 list that published sustainability reports over the period 2017-2023. We focus on this population because previous research demonstrates that biodiversity reporting remains rare beyond the world's largest corporations [1, 22]. To avoid noise from firms with sporadic reporting practices, we only considered those that published at least four reports in the analyzed period. Our final sample consists of 296 firms, and 1,783 firm-year observations.

We adopt the definitions of our two dependent variables from Bingler et al. [5], who developed these measures and applied them to climate-related disclosures:

The *Cheap Talk Index* (CTI) measures the proportion of non-specific commitments in corporate disclosures. It is defined as:

$$\mathrm{CTI}_{i,t} = \frac{\text{Number of non-specific commitment paragraphs}_{i,t}}{\text{Number of commitment paragraphs}_{i,t}}. \tag{1}$$

To construct the CTI, we first identified all biodiversity-related paragraphs in each sustainability report and classified them as either commitment or non-commitment statements. Commitment paragraphs were then filtered and classified as specific or non-specific. The CTI was subsequently computed for each firm-year observation.

The *Positive Talk Index* (PTI) captures the relative emphasis on positively framed biodiversity discourse in corporate reporting. It is defined as:

$$\mathrm{PTI}_{i,t} = \log\left(\frac{1 + \text{Number of positive paragraphs in report}_{i,t}}{1 + \text{Number of negative paragraphs in report}_{i,t}}\right). \tag{2}$$

To construct the PTI, biodiversity-related paragraphs were classified into positive, neutral, or negative. Subsequently, the index was calculated for each firm-year observation.

For both the CTI and PTI, we excluded firm-years with fewer than five relevant paragraphs to avoid instability and inflated ratios from observations presenting minimal biodiversity-related content. However, our results are robust to alternative thresholds, including more relaxed criteria, as presented in Section 6.

Our analysis incorporates two key independent variables:

Biodiversity Financial Risk Score - We measure firms' exposure to BRFR employing the sum of the physical and transition biodiversity risk scores developed by Giglio et al. [18]. These indicators capture the extent to which different industries face BRFR through both physical impacts (direct economic harm derived from the loss of biodiversity and ecosystem services) and transition risks (exposure to policy and market changes). The scores are available at the GICS 4-digit level. Their sum ranges from 1 to 48, with higher values indicating higher risk exposure. In our sample, the highest-scoring industries include: Energy, Food, Beverage & Tobacco, and Pharmaceuticals, Biotechnology & Life Sciences, while those with lower risk exposure include Software Services, Telecommunications Services, and Semiconductor Equipment.

Biodiversity Policy Stringency Score - We measure the strength of biodiversity protection policies in countries where firms are headquartered employing the Biodiversity & Habitat Index developed by Yale EPI [6]. This indicator "assesses countries' actions toward retaining natural ecosystems and protecting the full range of biodiversity within their borders."[16], and ranges 0-100, with higher scores indicating stronger biodiversity protection policies. The index is constructed from seven indicators encompassing the following dimensions: protection of terrestrial biomes; protection of marine biomes; state of conservation of natural habitats; quality of the protection afforded to terrestrial flora and fauna; and degree of habitat degradation; and representativeness of protected areas. Top-scoring countries in our sample include Austria, Germany, and Luxembourg, while the ones with the lowest scores are China, India, and Turkey.

We included the following controls in our regressions: ROA, size (logarithm of assets), leverage, asset growth, net income growth, LSEG-Refinitiv environmental pillar score, years listed in Fortune Global 500, report length (in pages), number of biodiversity-related paragraphs, and reporting frameworks followed (GRI, SASB, ISSB, ESRS, TCFD, TNFD). To account for differences in external verification practices, we additionally control for whether

reports receive external assurance, the type of assurer (big four vs. other providers), and the scope of assurance (limited vs. comprehensive). Because assurer type and assurance scope are only observed for firms that obtain assurance, including these variables separately would mechanically exclude non-assured firms. We therefore combine them into a single saturated firm-year "assurance profile" variable with mutually exclusive categories, using no assurance as the reference group. All regressions further include controls for reporting year, region, and industry.

Financial data was retrieved from S&P-Compustat, and ESG data from Refinitiv. Reporting frameworks were identified using a custom PyMuPDF-based text-extraction pipeline that parses each PDF, normalizes text, detects framework identifiers and proximity-based phrasing, and filters out negated or non-applicable mentions. This system was extensively validated to ensure that only genuine adoption of each framework was flagged. Assurance variables were manually coded by examining each report, as assurance statements are frequently provided through external links or scanned documents that cannot be reliably extracted automatically.

4.2. Model Specification

We utilize correlated random effects (CRE) models for this study. Traditional fixed effects estimation is unsuitable given that our key explanatory variables are constant in the period analyzed. The CRE approach [31, 42] addresses this constraint by controlling for potential correlation between unobserved firm-specific effects and the regressors, thereby allowing for the estimation of time-invariant variables.

We estimate two CRE models:

$$CTI_{it} = \beta X_{it} + \gamma \overline{X}_i + \delta Z_i + \alpha_i + \varepsilon_{it} \tag{3}$$

$$PTI_{it} = \beta X_{it} + \gamma \overline{X}_i + \delta Z_i + \alpha_i + \varepsilon_{it} \tag{4}$$

Where X is the vector of time-varying predictors, $\overline{X}$ is the vector of firm-level means of X (Mundlak correction), Z is the vector of time-invariant variables, α_i captures unobserved firm-specific effects, and ε_{it} is the idiosyncratic error term. Because reporting-framework and assurance-profile indicators exhibit low within-firm variation over time, their means would provide little identifying information while introducing near-collinearity issues. Following

standard practice in Mundlak specifications, we apply the correction only to covariates with substantial time variation.

Independent variables and controls were winsorized at the 1st and 99th percentiles to reduce the influence of outliers, with the exception of firm size. Firm size is included as the natural logarithm of total assets, and the log transformation already mitigates the influence of extreme values. The dependent variables were not winsorized, as extreme values reflect meaningful heterogeneity in biodiversity reporting practices. Multicollinearity was assessed using the condition number of the standardized design matrix (excluding fixed-effect dummies), yielding $\kappa = 26.6$, which indicates no cause for concern. We employed standard errors clustered at the firm level to account for potential heteroskedasticity and serial correlation within firms. Our results are robust to the use of firm-cluster wild bootstrap standard errors, as shown in Section 6.

4.3. Descriptive Statistics

Table 3 provides the number of firm-year observations by country and industry. The sample includes 1,783 observations, with a strong representation from the United States, Japan, and China. On the industry side, Financials, Energy, and Consumer Discretionary account for the largest shares, reflecting sectors with significant exposure to biodiversity-related risks and ESG scrutiny.

Table 3 – Number of firm-year observations by country and industry

Country	N	Industry	N
USA	597	Financials	394
Japan	211	Energy	242
China	186	Consumer discretionary	223
Germany	109	Consumer staples	191
France	107	Industrials	179
Canada	78	IT	173
Korea	75	Healthcare	138
UK	59	Materials	112
Switzerland	48	Communication services	71
Brazil	39	Diversified	60
India	37		
Italy	35		
Spain	28		
Taiwan	25		
Russia	24		
Ireland	14		
Netherlands	14		
Denmark	13		
Mexico	13		
Singapore	10		
Other	61		
Total	1783	Total	1783

Table 4 reports descriptive statistics for the variables used in the regression analysis. The Cheap Talk and Positive Talk Indices display substantial variation across firms, with means around 0.50 and 1.65, respectively[1]. Biodiversity Policy Stringency and Biodiversity Risk Score also exhibit wide dispersion, suggesting significant cross-country and cross-industry heterogeneity. Control variables such as ROA, Leverage, and Growth indicators show expected distributions for a sample of large global firms, with an average of 22 years in the Fortune Global 500, reflecting their maturity and market relevance.

[1]The N of the PTI and CTI are lower than that of the full sample, because not all reports contain biodiversity-related paragraphs. The N of the CTI is smaller than that of the PTI because the former requires ≥ 5 commitment paragraphs, while the latter requires ≥ 5 biodiversity-related paragraphs, including both commitments and non-commitments.

Table 4 – Descriptive statistics for regression variables

Variable	N	mean	std	min	10%	50%	90%	max
Cheap Talk Index	1499	0.50	0.30	0.00	0.00	0.50	1.00	1.00
Positive Talk Index	1550	1.65	0.87	-0.69	0.69	1.61	2.80	3.60
Biodiv Policy Stringency Score	1783	44.77	17.78	11.00	11.00	41.00	70.00	82.00
Biodiv Financial Risk Score	1783	27.97	14.65	2.00	2.00	27.00	46.00	47.00
Environmental Score	1727	75.43	16.40	22.02	52.55	78.56	94.12	97.89
Log Assets	1750	11.71	1.30	9.20	10.25	11.48	13.75	14.91
ROA	1745	0.04	0.05	-0.08	0.00	0.03	0.11	0.24
Leverage	1745	0.26	0.15	0.00	0.10	0.20	0.50	0.70
Net Income Growth	1717	0.14	2.46	-10.38	-1.04	0.03	1.23	14.62
Assets Growth	1715	0.06	0.14	-0.22	-0.07	0.04	0.20	0.77
Num. Biodiversity Paragraphs	1783	15	22	0	0	6	45	114
Number of Pages	1783	118	85	12	37	93	242	415
Years in FG500	1783	22	9	1	7	26	30	30

Figure 2 shows the evolution of the average Cheap Talk Index (CTI) from 2017 to 2023 for firm-year observations with at least five commitment paragraphs. The CTI exhibits a clear upward trend over time, with particularly marked increases from FY 2020 onward, suggesting an increasing reliance on vague biodiversity-related commitments in corporate reporting.

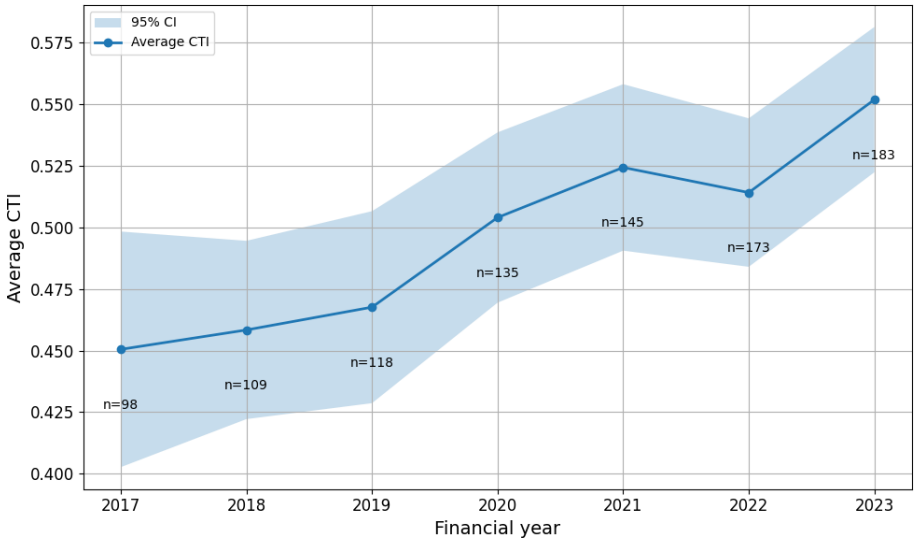


Figure 2: Average CTI by Year

Figure 3 displays the yearly average of the positive-to-negative sentiment ratio in biodiversity reporting (before log-transformation), for firm-year observations with at least five biodiversity-related paragraphs. The ratio also trends upward, especially after FY 2020, indicating a consistent shift toward more optimistic framing in corporate biodiversity discourse.

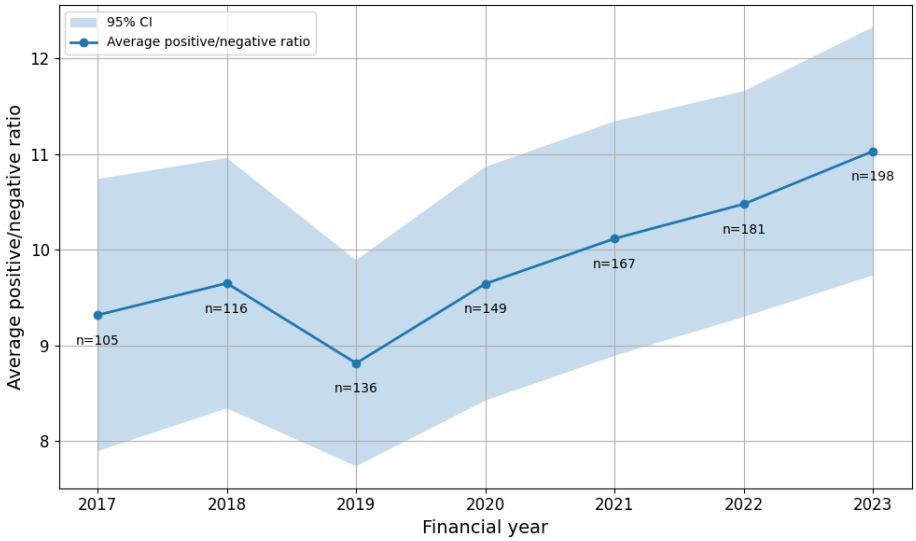


Figure 3: Average positive-to-negative sentiment by Year

Together, these figures suggest that both forms of greenwashing—cheap talk and positive talk—have intensified over time, particularly after the key institutional milestones of 2021 and 2022 consistent with H3.

5. Results

5.1. Main regression results

Table 5 reports the CRE regression results for the CTI. We estimate three specifications that vary in their fixed effects structure to ensure robustness: Model (1) includes year, industry, and region fixed effects, Model (2) industry-by-year and region fixed effects, and Model (3) instead combines industry and region-by-year fixed effects.

The findings across columns (1)–(3) provide strong support for H1, showing that exposure to BRFR is positively associated with cheap talk. In all

specifications, the coefficient on the Biodiversity Financial Risk Score remains positive and highly statistically significant $p < 0.01$. For instance, the estimated coefficient in Model (2) is 0.0087, implying that a one standard deviation increase in the score (≈ 14.65 points) translates into a 12.7 percentage-point increase in the CTI. Expressed in percentile terms, moving from the 25th to the 75th percentile of biodiversity-related financial risk exposure (an interquartile increase of 25 points) is associated with a 22 percentage-point increase in the CTI. This suggests that BRFR-exposed firms rely more heavily on non-specific commitments. This is consistent with the interpretation of cheap talk as a symbolic disclosure strategy to manage legitimacy pressures while preserving strategic flexibility. The temporal dummies further support H3. Cheap talk increases significantly from reports published in 2021 (FY 2020) onwards, suggesting that firms intensified their use of vague biodiversity-related commitments around the Kunming–Montreal process. This pattern is consistent with anticipatory disclosure behavior, whereby firms expand symbolic commitments in response to rising institutional salience prior to the formal adoption of international biodiversity agreements.

Table 5 – CRE Cheap Talk Results

Variables	(1)	(2)	(3)
Biodiversity Policy Stringency Score	0.0013 (0.0011)	0.0013 (0.0012)	0.0011 (0.0011)
Biodiversity Financial Risk Score	0.0072*** (0.0021)	0.0087*** (0.0025)	0.0070*** (0.0021)
Environmental Score	-0.001 (0.0011)	-0.0004 (0.0013)	-0.0007 (0.0012)
log Assets	0.0146 (0.0685)	0.073 (0.0791)	0.0307 (0.0719)
ROA	-0.1685 (0.2086)	-0.033 (0.2124)	-0.2074 (0.2233)
Leverage	-0.4832*** (0.1459)	-0.3398** (0.1418)	-0.5102*** (0.1492)
Years in FG500	0.0001 (0.0015)	0.0005 (0.0016)	-0.0 (0.0015)
Assets Growth	0.0334 (0.0567)	0.0051 (0.0653)	0.0287 (0.0598)
Net Income Growth	-0.0016 (0.0013)	-0.0014 (0.0017)	-0.0014 (0.0013)
Num. Biodiversity Paragraphs	-0.0008* (0.0004)	-0.0007* (0.0004)	-0.0009* (0.0005)
Number of Pages	0.0002 (0.0001)	0.0001 (0.0002)	0.0002 (0.0002)
GRI	-0.0173 (0.0181)	-0.0255 (0.02)	-0.0158 (0.0182)
SASB	0.0261 (0.0189)	0.0387** (0.0187)	0.0157 (0.0203)
ISSB	0.0339 (0.0268)	0.0323 (0.0266)	0.0323 (0.029)
ESRS	0.0117 (0.0353)	0.0177 (0.0396)	0.0194 (0.038)
TCFD	0.0038 (0.017)	0.0059 (0.0171)	0.0122 (0.0177)
TNFD	-0.0254 (0.0391)	-0.008 (0.0406)	-0.0179 (0.0402)
Assurance Profile Big 4 Comprehensive	-0.0057 (0.0227)	-0.0168 (0.0265)	-0.0078 (0.0231)
Assurance Profile Big 4 Limited	0.0091 (0.0432)	-0.0092 (0.0466)	0.0094 (0.0457)
Assurance Profile Other Comprehensive	0.0554 (0.0719)	0.066 (0.0693)	0.0375 (0.0723)
Assurance Profile Other Limited	-0.0622** (0.0316)	-0.0852** (0.0411)	-0.0684** (0.0328)
Year 2018	0.0232 (0.0237)		
Year 2019	0.0373 (0.0244)		
Year 2020	0.0819*** (0.0296)		
Year 2021	0.0654** (0.0298)		
Year 2022	0.066* (0.0353)		
Year 2023	0.0996*** (0.0336)		
mean Environmental Score	0.0031** (0.0014)	0.0027* (0.0015)	0.0029** (0.0014)
mean log Assets	0.003 (0.0674)	-0.0531 (0.0781)	-0.0096 (0.0711)
mean ROA	0.655* (0.3626)	0.6446* (0.3762)	0.7293** (0.3711)
mean Leverage	0.4696*** (0.1678)	0.3441** (0.1661)	0.5051*** (0.1698)
mean Assets Growth	-0.1959 (0.2171)	-0.159 (0.2314)	-0.1985 (0.22)
mean Net Income Growth	-0.004 (0.0073)	-0.0036 (0.0078)	-0.0044 (0.0072)
mean Num. Biodiversity Paragraphs	0.0001 (0.0007)	-0.0002 (0.0007)	0.0002 (0.0007)
mean Number of Pages	-0.0001 (0.0002)	0.0001 (0.0002)	-0.0001 (0.0002)
FE industry	YES	NO	YES
FE industry × year	NO	YES	NO
FE region	YES	YES	NO
FE region × year	NO	NO	YES
R-squared (overall)	0.88	0.89	0.88
R-squared (between)	0.92	0.92	0.92
R-squared (within)	0.07	0.18	0.09

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Obs: 896

Among the control variables, leverage presents an interesting dual effect.

Within-firm increases are associated with lower cheap talk, suggesting that as firms become more leveraged, they face constraints that limit such behavior. Conversely, firms with persistently higher leverage levels display greater reliance on cheap talk, indicating that financial structure influences the extent to which firms resort to vague biodiversity-related commitments.

Turning to the PTI results, we find strong support for H2. As presented in Table 6, the Biodiversity Policy Stringency Score is positively and highly significant across all specifications ($p < 0.01$). The estimated coefficient on the Biodiversity Policy Stringency Score in the PTI regression is 0.0089 (log-ratio units) in Models (1) and (3). A one standard deviation increase in policy stringency (≈ 17.8 points) implies a 0.158 rise in the log-ratio of positive to negative biodiversity-related paragraphs, which corresponds to roughly a 17.2% increase in the ratio. Expressed in percentile terms, moving from the 25th to the 75th percentile of biodiversity policy stringency (an interquartile increase of 19 points) is associated with an increase of approximately 0.17 in the log ratio, corresponding to about an 18.4% increase in the positive-to-negative biodiversity disclosure ratio. This effect is large, suggesting it could meaningfully alter the stakeholders' perceptions of biodiversity-related risks.

Table 6 – PTI CRE Results

Variables	(1)	(2)	(3)
Biodiversity Policy Stringency Score	0.0089*** (0.003)	0.0085*** (0.003)	0.0089*** (0.003)
Biodiversity Financial Risk Score	0.0064 (0.0103)	0.0051 (0.0109)	0.0063 (0.0104)
Environmental Score	-0.0032 (0.0026)	-0.0019 (0.0028)	-0.0029 (0.0027)
log Assets	-0.1059 (0.1501)	-0.0935 (0.1779)	-0.1111 (0.1546)
ROA	-0.7598 (0.5641)	-0.3239 (0.5785)	-0.5203 (0.5598)
Leverage	0.311 (0.4141)	0.4201 (0.4775)	0.4126 (0.4383)
Years in FG500	-0.0062 (0.0044)	-0.0061 (0.0045)	-0.0056 (0.0044)
Assets Growth	-0.0674 (0.1867)	-0.0538 (0.2127)	-0.1209 (0.2081)
Net Income Growth	0.011 (0.0082)	0.0115 (0.0083)	0.011 (0.0084)
Num. Biodiversity Paragraphs	0.0105*** (0.0018)	0.0095*** (0.0018)	0.0109*** (0.0019)
Number of Pages	-0.0001 (0.0005)	-0.0001 (0.0005)	-0.0002 (0.0005)
GRI	-0.0285 (0.0556)	-0.0272 (0.0575)	-0.0254 (0.0567)
SASB	-0.1507** (0.0619)	-0.1584** (0.0636)	-0.1512** (0.0645)
ISSB	-0.1193 (0.0883)	-0.0837 (0.0937)	-0.1153 (0.0902)
ESRS	-0.1467 (0.1192)	-0.1253 (0.1188)	-0.0858 (0.1254)
TCFD	0.0691 (0.0494)	0.0633 (0.0529)	0.0735 (0.0504)
TNFD	-0.0488 (0.1123)	-0.02 (0.1055)	-0.0647 (0.1206)
Assurance Profile Big 4 Comprehensive	-0.029 (0.0757)	-0.0513 (0.0746)	-0.0226 (0.0766)
Assurance Profile Big 4 Limited	-0.1133 (0.0839)	-0.1594* (0.0906)	-0.111 (0.086)
Assurance Profile Other Comprehensive	-0.1882 (0.1244)	-0.2545* (0.1416)	-0.1744 (0.1268)
Assurance Profile Other Limited	0.1061 (0.1682)	0.1056 (0.1777)	0.1184 (0.1752)
Year 2018	0.0869 (0.0618)		
Year 2019	-0.0266 (0.0745)		
Year 2020	0.0508 (0.0875)		
Year 2021	0.1273 (0.0848)		
Year 2022	0.2439*** (0.0846)		
Year 2023	0.2660*** (0.0865)		
mean Net Income Growth	-0.0054 (0.0303)	-0.0003 (0.0308)	-0.0051 (0.0307)
mean Environmental Score	0.0075** (0.0037)	0.0056 (0.0039)	0.0074* (0.0038)
mean ROA	2.5503** (1.179)	1.9532* (1.1861)	2.2767* (1.1934)
mean Num. Biodiversity Paragraphs	0.0021 (0.0024)	0.0033 (0.0024)	0.0017 (0.0025)
mean Number of Pages	0.0008 (0.0007)	0.0008 (0.0007)	0.0009 (0.0008)
mean Leverage	-0.1904 (0.5067)	-0.2975 (0.5557)	-0.3065 (0.529)
mean Assets Growth	0.5394 (0.5681)	0.4499 (0.5665)	0.5753 (0.5864)
mean log Assets	0.193 (0.1482)	0.1646 (0.1761)	0.1944 (0.1533)
FE industry	YES	NO	YES
FE industry × year	NO	YES	NO
FE region	YES	YES	NO
FE region × year	NO	NO	YES
R-squared (overall)	0.91	0.92	0.91
R-squared (between)	0.94	0.95	0.94
R-squared (within)	0.16	0.25	0.17

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Obs: 982

Temporal dummies also reveal that positive sentiment intensified markedly after FY 2021, consistent with H3. This pattern aligns with firms adopting a more optimistic framing in their biodiversity-related disclosures in response to heightened legitimacy pressures following the adoption of the Kunming Declaration and the Global Biodiversity Framework. Among control variables, following SASB reporting standards is negatively associated with the PTI. Conversely, the mean firm profitability (ROA) is positively related to sentiment, implying that persistently more profitable firms are particularly prone to emphasizing the positive aspects of their biodiversity discourse.

Comparing the results from Tables 5 and 6, we observe two distinct patterns in biodiversity-related greenwashing. Higher exposure to biodiversity-related financial risks is associated with higher levels of cheap talk (supporting H1). This pattern is consistent with the use of non-specific disclosures to convey risk assurance while avoiding substantive costs and retaining strategic flexibility. In contrast, firms headquartered in countries with more stringent biodiversity protection policies exhibit systematically higher levels of positive talk (supporting H2). This pattern is consistent with an emphasis on favorable biodiversity-related narratives to signal alignment with regulatory and societal expectations. Taken together, these findings suggest that distinct sources of legitimacy pressure are associated with different forms of symbolic biodiversity disclosure: financial risk exposure correlates with non-specific commitments, while regulatory and societal pressure correlate with more optimistic framing of biodiversity-related content.

Importantly, both cheap talk and positive talk intensified around the Kunming–Montreal period (supporting H3), with cheap talk increasing earlier and positive talk strengthening more clearly after the consolidation of international biodiversity frameworks. This pattern is consistent with an initial reliance on vague commitments under rising investor scrutiny, followed by a broader tonal adjustment as biodiversity expectations became more widely institutionalized.

Our results remain stable across a broad set of robustness checks, including alternative model specifications, regional–sectoral fixed effects, bootstrap inference, and alternative constructions of the disclosure measures, as detailed in Section 6.

To assess the explanatory contribution of our key regressors, we compute between-firm partial R^2 measures from firm-level cross-sectional regressions. Since BRFR is defined at the industry level and policy stringency at the country level, both variables are time-invariant for individual firms in our

panel. Their coefficients in the CRE–Mundlak model are therefore identified exclusively from between-firm variation. Accordingly, partial R^2 statistics are calculated by regressing firm-level means of the dependent variable on firm-level averages of time-varying controls, time-invariant characteristics, and industry and region fixed effects.

The partial R^2 then captures the incremental proportion of between-firm variance explained by each key variable beyond these controls. BRFR explains 4.07% of the between-firm variation in the CTI, whereas policy stringency explains 1.64% of the between-firm variation in the PTI. The smaller contribution of policy stringency reflects its coarser level of variation (country vs. industry), yet both variables contribute meaningfully to cross-sectional disclosure heterogeneity. We interpret these partial R^2 values as descriptive measures of explanatory contribution conditional on our controls, not as causal effect sizes.

5.2. Dynamic patterns around Kunming–Montreal: Event-study analysis

While the baseline CRE Mundlak specifications document a marked increase in both cheap talk and positive talk, year dummies alone cannot isolate biodiversity-specific dynamics from broader shifts in sustainability reporting. To clarify the timing of disclosure changes and assess pre-event dynamics, we therefore complement the baseline estimates with an event-study design. The event of interest is the adoption of the Kunming Declaration in October 2021, with event time mapped to financial reporting years to reflect when disclosure decisions become observable.

Event time is indexed using financial reporting year indicators, with FY 2019 serving as the omitted reference category. FY 2021 corresponds to the first set of sustainability reports published after the Kunming declaration. Importantly, FY corresponds to the reporting year covered by sustainability reports, which are typically published in the following calendar year. Accordingly, FY 2021 reports are released in 2022 and capture disclosure decisions made after the Kunming Declaration.

We estimate event-study specifications within the CRE Mundlak framework, replacing FY dummies with a set of lead and lag indicators indexed to event time. The model takes the form:

$$y_{it} = \sum_{t \neq 2019} \beta_t \, \mathbf{1}[\mathrm{FY}_{it} = t] + \sum_{t \neq 2019} \delta_t \, \mathbf{1}[\mathrm{FY}_{it} = t] \times EU_i + X'_{it}\gamma + \bar{X}'_i\theta + Z'_i\lambda + u_i + \varepsilon_{it}. \tag{5}$$

where y_{it} denotes either the CTI or PTI for firm i in year t. Financial reporting year indicators capture dynamic disclosure patterns over time, with FY 2019 serving as the omitted baseline category. X_{it} includes time-varying firm-level controls, $\bar{X}_i$ their firm-specific means (Mundlak terms), and Z_i time-invariant controls, including region, industry, and assurance profiles. u_i denotes the firm-specific random effect and ε_{it} the idiosyncratic error term. Standard errors are clustered at the firm level. The interaction terms with EU_i allow for differential dynamics between EU and non-EU firms, with coefficients δ_t capturing EU-specific deviations from the baseline trend.

This specification allows us to test for parallel pre-trends and to trace the dynamic evolution of disclosure behavior in the period leading up to and following the Kunming COP.

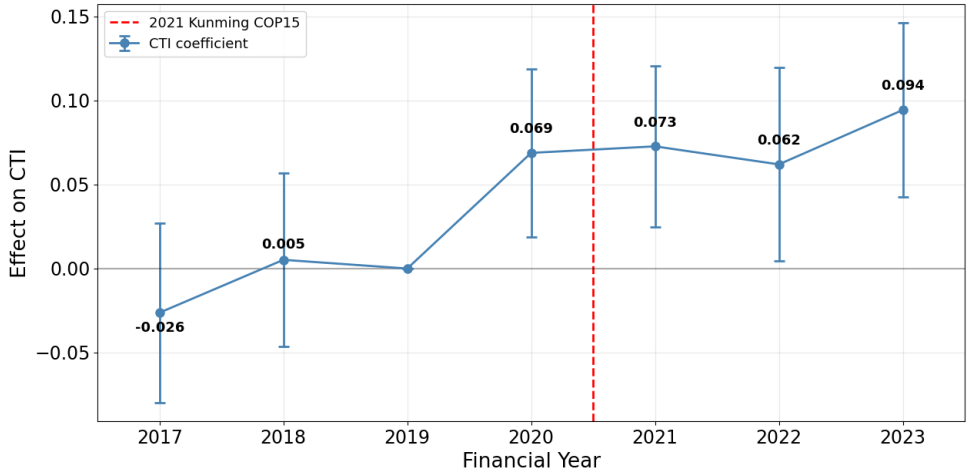


Figure 4: Event study - CTI Coefficients relative to FY 2019 baseline

Figure 4 plots the estimated event-time coefficients for cheap talk relative to the FY 2019 baseline, with point estimates and 95% confidence intervals. Formal tests fail to reject the null hypothesis of parallel pre-trends: a joint Wald test of pre-baseline coefficients yields $\chi^2(2) = 1.56$ ($p = 0.46$), and individual lead estimates are small and statistically insignificant. This supports the validity of the event-study design.

In contrast, post-FY 2020 dynamics display a clear and sustained increase in cheap talk. Estimated effects turn positive for FY 2020 reports and remain statistically significant throughout the post-Kunming period, with magnitudes increasing over time. This pattern aligns with a pre-emptive

expansion of vague biodiversity-related commitments, becoming visible in reports released around the Kunming negotiations[2] and strengthening thereafter, consistent with rising legitimacy pressures and anticipatory disclosure behavior.

EU interaction terms suggest some heterogeneity in these dynamics. The interaction for FY 2021 is negative and statistically significant ($\delta_{2021} = -0.141$, $p = 0.016$), indicating that EU firms exhibit a weaker contemporaneous increase in cheap talk relative to non-EU firms. A joint test of post-2021 EU interaction terms provides marginal evidence of differential dynamics ($\chi^2(3) = 6.69$, $p = 0.08$). While these effects should be interpreted cautiously, they are consistent with EU firms facing stronger regulatory scrutiny that partially constrains the expansion of symbolic biodiversity discourse.

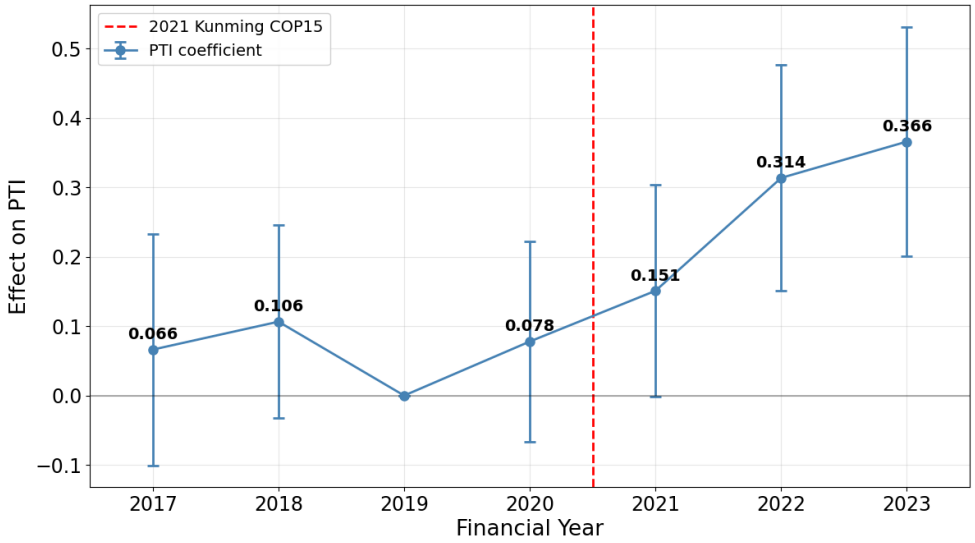


Figure 5: Event study - PTI Coefficients relative to FY 2019 baseline

Figure 5 reports the corresponding event-study coefficients for positive talk, formatted identically to Figure 4. Parallel-trends tests again provide no evidence of differential dynamics in earlier years: a joint test of pre-baseline

[2]FY 2020 sustainability reports were published in calendar year 2021, predominantly prior to the October 2021 Kunming Declaration. This timing places these disclosures in the anticipatory window preceding the COP, when heightened biodiversity scrutiny was foreseeable but before formal adoption of the declaration.

coefficients yields $\chi^2(2) = 2.27$ ($p = 0.32$), and individual lead estimates are statistically insignificant.

Post-event effects for positive talk emerge later than for cheap talk. Coefficients become positive in FY 2021 but are only weakly significant, while strong and statistically significant increases are observed from FY 2022 onwards. This delayed response suggests that positive talk intensifies once biodiversity expectations become more broadly institutionalized.

EU interaction terms for positive talk are individually imprecisely estimated but consistently negative in the post-event period. A joint test of post-FY 2021 EU interaction terms yields weak evidence of differential dynamics ($\chi^2(3) = 6.49$, $p = 0.09$), suggesting that stricter regulatory environments may attenuate the expansion of optimistic biodiversity framing.

Taken together, the event-study results provide strong support for H3 while revealing three key insights into the temporal dynamics of biodiversity greenwashing. First, the increase in cheap talk begins earlier (from FY 2020) and builds more gradually than the rise in positive talk. Second, positive talk intensifies later (from FY 2022) but with substantially larger coefficients. Third, EU firms display weaker post-event increases in both forms of symbolic disclosure, although the evidence is only marginally significant for both cheap talk ($p = 0.08$) and positive talk ($p = 0.09$).

This staggered timing is consistent with the mechanisms outlined in Section 2. As BRFR gained salience among investors in the period leading up to the Kunming Declaration, firms initially expanded cheap talk, using vague and non-verifiable commitments as a low-cost strategy to reassure investors under conditions of high uncertainty and limited accountability. Following the Kunming process, as biodiversity expectations diffused beyond financial markets to a broader set of stakeholders, firms increasingly complemented cheap talk with positive talk, emphasizing favorable biodiversity narratives to signal alignment with emerging societal norms.

While the event-study design is not intended to establish causal policy effects, it provides a transparent characterisation of the timing of disclosure adjustments around the Kunming milestone. The FY 2019 baseline satisfies parallel-trends assumptions for both outcomes, lending credibility to a descriptive interpretation of post-event disclosure dynamics.

Robustness checks using FY 2020 as the baseline are reported in the Appendix H and confirm that this reference period violates parallel trends for cheap talk, reinforcing the choice of FY 2019 as the preferred baseline.

6. Robustness

6.1. Alternative Model Specifications

To assess whether our findings are sensitive to model specification, we re-estimate our main regressions using pooled OLS with standard errors clustered at the firm level. Although this specification does not absorb unobserved firm heterogeneity, it provides a transparent benchmark against which to compare our preferred CRE models. The full pooled OLS regression results are reported in Appendix H. Across all specifications, the coefficients of our key explanatory variables remain highly statistically significant, with magnitudes very similar to those obtained in the preferred CRE models. These results indicate that the associations between BRFR exposure and cheap talk, and between policy stringency and positive talk, are not driven by the correlated random effects structure.

For the Cheap Talk Index, the pooled OLS results reported in Table A.23 confirm that BRFR exposure is strongly and positively associated with cheap talk across all specifications. The estimated coefficients (0.0062–0.0071) are only slightly smaller than those obtained under the CRE models (0.0072–0.0087), reinforcing H1. However, the temporal pattern is less pronounced in OLS: relative to 2017, the year dummies are generally positive but imprecisely estimated, and only the 2023 dummy is statistically significant. This contrasts with the CRE models, where cheap talk is significantly higher in reports published from 2021 onwards, and highlights the importance of accounting for unobserved firm heterogeneity when identifying temporal shifts in disclosure practices.

Appendix Table A.24 reports the pooled OLS estimates for the Positive Talk Index. The results confirm that biodiversity policy stringency remains positively and significantly associated with positive talk (coefficients ranging 0.0061–0.0064) under the OLS specification, supporting H2 and closely matching the magnitudes obtained in the CRE models. In contrast, none of the year dummies is statistically significant in OLS, whereas the CRE results show a marked increase in positive talk in FY 2022–2023. This again suggests that controlling for unobserved firm-level heterogeneity is essential for cleanly identifying temporal dynamics in disclosure sentiment.

6.2. Region × Industry Fixed Effects

A potential concern is that the observed associations may be driven by systematic differences across geographical–sectoral clusters rather than firm-

level dynamics. To address this, we re-estimated our CRE Mundlak models including a full set of region × industry fixed effects. This saturated structure absorbs all time-invariant heterogeneity common to firms operating within the same regional–sectoral context, thereby tightening identification and providing a demanding robustness test.

For the Cheap Talk Index (Table 7), the coefficient on the BRFR Score remains positive, stable in magnitude, and statistically significant at the 1% level (0.0077; SE = 0.0027). The post-2020 year effects also retain their sign, size, and significance, closely mirroring the baseline specification. Model fit measures remain largely unchanged, indicating that the observed rise in cheap talk is not driven by compositional patterns across region–industry clusters.

Table 7 – CRE CTI Results with Region × Industry FE

Variable	Coefficient	Clustered SE
Biodiversity Financial Risk Score	0.0077***	(0.0027)
Year 2020	0.0839***	(0.0303)
Year 2021	0.0664**	(0.0304)
Year 2022	0.0680*	(0.0359)
Year 2023	0.1023***	(0.0343)
R-squared (between)	0.93	
R-squared (within)	0.07	
R-squared (overall)	0.88	

Notes: *** p<0.01, ** p<0.05, * p<0.1. Obs:896

For the Positive Talk Index (Table 8), the coefficient on the Biodiversity Policy Stringency Score remains positive, statistically significant, and nearly identical in magnitude to the baseline CRE estimates (0.0087; SE = 0.0030). The increases in positive talk observed after the adoption of the Kunming–Montreal framework also persist, with the FY 2022 and FY 2023 coefficients (0.247 and 0.275, respectively) closely matching those reported in the main analysis. Once again, model fit measures remain stable. Taken together, these results demonstrate that both the BRFR–cheap talk association and the policy stringency–positive talk association are robust to controlling for all fixed region–industry characteristics, reinforcing the validity of our main findings.

Table 8 – CRE PTI Results with Region × Industry FE

Variable	Coefficient	Clustered SE
Biodiversity Policy Stringency Score	0.0087***	(0.0030)
Year 2022	0.2472***	(0.0867)
Year 2023	0.2746***	(0.0880)
R-squared (between)	0.95	
R-squared (within)	0.16	
R-squared (overall)	0.92	

Notes: *** p<0.01, ** p<0.05, * p<0.1. Obs:982

6.3. Alternative inference procedures

We re-assess statistical significance using the wild cluster bootstrap, which is widely regarded as a reliable inference procedure for settings with a moderate number of clusters and unbalanced panels. Table 9 compares the original CRE estimates with wild cluster bootstrap standard errors (1,999 replications, Rademacher weights). The coefficient on the BRFR score remains positive and highly significant (0.0072, p<0.01), with standard errors declining 48% (0.0021 → 0.0011). Year effects retain their sign and magnitude, with standard errors declining an average of 14% across year dummies. Importantly, no coefficient loses significance and no signs change.

Table 9 – CRE CTI results with Clustered and Wild Bootstrap SE

Variable	Coefficient (CRE)	Cluster SE	Wild bootstrap SE
Biodiversity Financial Risk Score	0.0072***	0.0021	0.0011
Year 2020	0.0819***	0.0296	0.0263
Year 2021	0.0654**	0.0298	0.0259
Year 2022	0.0660**	0.0353	0.0293
Year 2023	0.0996***	0.0336	0.0277

Notes: Coefficients come from Model (1) of the CTI CRE Mundlak specification in Table 5. 896 Obs. 213 Clusters (firms). *** p<0.01, ** p<0.05, * p<0.1.(Wild bootstrap)

Table 10 presents the same robustness exercise for the PTI. The coefficient on the Biodiversity Policy Stringency Score remains positive and strongly significant (0.0089, p<0.01), with bootstrap standard errors 47% smaller (0.0030 → 0.0016). The reduction in standard errors is consistent with the bootstrap

correcting finite-sample overestimation in clustered panels. The 2022 and 2023 increases remain statistically significant with 10–12% smaller standard errors. The 2018-2021 dummies remain insignificant under both variance estimators and are omitted for parsimony.

Table 10 – CRE PTI Results with Clustered and Wild Bootstrap SE

Variable	Coefficient (CRE)	Cluster SE	Wild bootstrap SE
Biodiversity Policy Stringency Score	0.0089***	0.0030	0.0016
Year 2022	0.2439***	0.0846	0.0756
Year 2023	0.2660***	0.0865	0.0760

Notes: Coefficients come from Model (1) of the PTI Mundlak specification in Table 6. 982 Obs. 222 Clusters (firms). *** p<0.01, ** p<0.05, * p<0.1 (Wild bootstrap).

In sum, for both greenwashing practices examined in this study, the wild cluster bootstrap confirms that our core inferences are not sensitive to the choice of variance estimator, providing strong reassurance regarding the stability and reliability of the results. Moreover, the substantial reductions in standard errors under bootstrap inference suggest that our main specifications may actually provide conservative estimates of statistical significance, further bolstering confidence in the reported associations between biodiversity risk exposure, policy stringency, and corporate greenwashing behavior.

6.4. Alternative minimum disclosure thresholds

To test whether our results are sensitive to the minimum disclosure volume required to compute the Cheap Talk Index, we re-estimate the regressions using different thresholds for the minimum number of biodiversity-related commitment paragraphs (≥ 4 – ≥ 10). As reported in Table 11, the coefficients on BRFR exposure remain positive and highly significant at all thresholds, with similar magnitudes. This indicates that the link between BRFR exposure and cheap talk is not an artifact of the chosen cutoff.

Table 11 - Robustness to Alternative Paragraph Thresholds - CTI

Threshold	BRFR Coefficient	p-value	Std. Error	Observations	R^2
≥ 4	0.0069***	0.004	(0.0024)	1005	0.86
≥ 5	0.0087***	0.006	(0.0025)	896	0.88
≥ 6	0.0073***	0.003	(0.0024)	840	0.89
≥ 7	0.0073***	0.005	(0.0026)	757	0.90
≥ 8	0.0079***	0.004	(0.0027)	714	0.90
≥ 9	0.0074***	0.009	(0.0028)	653	0.90
≥ 10	0.0059**	0.012	(0.0023)	612	0.91

Notes: Results from Model 2 – Industry × Year and Region FE. Equivalent results are obtained for Models 1 and 3.

We also re-estimate the Positive Talk regressions using alternative paragraph thresholds (≥ 4 – ≥ 12). Results presented in Table 12 show that the coefficient on biodiversity policy stringency remains positive and statistically significant at the 1% level across all thresholds. The magnitude of the effect even increases slightly at higher thresholds, suggesting that the relationship between policy stringency and positive talk is particularly strong among firms with more extensive biodiversity reporting.

Table 12 - Robustness to Alternative Paragraph Thresholds - PTI

Threshold	Policy Stringency Coeff.	p-value	Std. Error	Observations	R^2
≥ 4	0.0092***	0.002	(0.0031)	1078	0.91
≥ 5	0.0085***	0.004	(0.0030)	982	0.92
≥ 6	0.0098***	0.003	(0.0033)	919	0.92
≥ 7	0.0109***	0.002	(0.0036)	833	0.93
≥ 8	0.0114***	0.003	(0.0038)	794	0.93
≥ 9	0.0117***	0.003	(0.0039)	729	0.93
≥ 10	0.0116***	0.004	(0.0040)	700	0.93
≥ 11	0.0130***	0.002	(0.0042)	640	0.94
≥ 12	0.0142***	0.002	(0.0045)	609	0.94

Notes: Results from Model 2 – Industry × Year and Region FE. Equivalent results are obtained for Models 1 and 3.

6.5. Alternative constructions of the PTI

We assessed the sensitivity of our results to the choice of smoothing constant in the PTI log-ratio transformation (Equation 2). Although adding a constant of 1 to the numerator and denominator is a standard practice to avoid division by zero and to moderate extreme ratios, it could be viewed as arbitrary. To address this, we perform a sensitivity analysis in which we vary the smoothing constant from 0.5 to 2.0 and re-estimate our main CRE-PTI specification for each value. Table 13 summarizes the results. The coefficient on biodiversity policy stringency remains positive and highly significant across all specifications, with magnitudes ranging from 0.0066 to 0.0105. As expected, the coefficient declines mechanically as the smoothing constant increases, because larger constants compress variation in the log-ratio transformation and therefore dampen the estimated response of PTI to changes in policy stringency. Model fit remains high (R^2 between 0.91 and 0.93), and standard errors decrease slightly as the constant increases, suggesting that extreme values exert limited influence on the estimated relationship.

Table 13 - Robustness to Alternative PTI Smoothing Constant

Constant	Policy Stringency Coeff.	p-value	Std. Error	R^2
0.5	0.0105***	0.0052	(0.0037)	0.91
0.75	0.0094***	0.0048	(0.0033)	0.92
1	0.0085***	0.0049	(0.0030)	0.92
1.5	0.0074***	0.0055	(0.0027)	0.92
2	0.0066***	0.0065	(0.0024)	0.93

Notes: Results from Model 2 – Industry × Year and Region FE. Equivalent results are obtained for Models 1 and 3. Obs: 982

6.6. Testing BRFR–post-2021 interactions

We re-estimated the CTI CRE Model (1) including an interaction between the BRFR score and a post-2021 indicator to examine whether industries with higher biodiversity-related financial risk disproportionately contributed to the post-Kunming increase in cheap talk. The interaction term is extremely small (0.0001) and statistically insignificant (p = 0.87). The BRFR effect remains positive and highly significant in both periods: 0.00697 (p = 0.0018) for reports published before 2021 and 0.00711 (p = 0.0014) for reports published from 2021 onwards. A Wald test confirms that these slopes

do not differ statistically ($\chi^2(1) = 0.03$, p = 0.87). These results indicate that the rise in cheap talk after 2021 was broad-based rather than driven by high-risk industries.

We also estimated the PTI CRE Model (1) including an interaction between BRFR and a post-2021 indicator to assess whether high-risk industries were disproportionately responsible for the rise in positive talk after 2021. Consistent with the baseline PTI results, BRFR is not a statistically significant predictor of positive talk in either period. The interaction term is small (-0.0035) and statistically insignificant (p = 0.22). The marginal effect of BRFR on PTI is weak and statistically indistinguishable across periods: 0.0082 (p = 0.42) for reports published before 2021 and 0.0047 (p = 0.65) for those published from 2021 onwards. A Wald test confirms that these slopes do not differ ($\chi^2(1) = 1.49$, p = 0.22). These results indicate that the post-2021 increase in positive talk was broad-based rather than concentrated in industries facing higher biodiversity-related financial risk.

6.7. Construct Validation of the Cheap Talk Index

To further assess the validity of the Cheap Talk Index, we conducted a targeted report-level audit of firm-year observations in the top decile of the CTI distribution. Manual evaluation confirms that 89% of paragraphs classified as non-specific commitments are genuinely vague, with the remaining cases reflecting boundary ambiguities rather than systematic classification error. A detailed account of this validation exercise is provided in Appendix J.

6.8. Sensitivity to extreme index values

Results are robust to alternative treatments of extreme values. Appendix F demonstrates that the main findings remain stable when winsorizing the dependent variables at progressively stricter thresholds (2.5–97.5% and 5–95%). Complementing this analysis, Appendix E addresses a related concern arising from the ratio-based construction of the Positive Talk Index. Because the PTI is computed as a log ratio, firm–year observations with few or no negative paragraphs can mechanically yield high index values. Re-estimating the models after restricting the sample to observations with more substantive negative disclosure shows that the coefficient on biodiversity policy stringency remains positive and statistically significant—and increases in magnitude—indicating that the results reflect systematic disclosure patterns rather than ratio-driven artifacts.

6.9. Limitations and identification considerations

While our correlated random effects approach allows us to estimate the association between time-invariant variables and greenwashing practices, methodological limitations prevent us from making strong causal claims.

First, our key explanatory variables—BRFR score and Biodiversity Policy Stringency Score—operate at the industry and country level, respectively, rather than varying at the firm level. This means they may capture broader industry or country characteristics beyond risk exposure or regulatory stringency that independently affect disclosure style. For BRFR, unobserved industry factors such as supply chain complexity, reporting standardization norms, litigation risk, or sector-specific stakeholder scrutiny could drive the observed associations. For policy stringency, country-level confounders such as corporate governance culture, media scrutiny, NGO pressure, legal traditions, or general environmental awareness could explain the positive talk patterns. Although we control for industry and region fixed effects and a range of firm characteristics, unobserved factors at the industry or country level correlated with both our explanatory variables and greenwashing practices cannot be fully ruled out.

Second, the time-invariant nature of both scores limits our ability to exploit within-firm variation that would strengthen identification. These limitations suggest our findings should be interpreted as robust associations rather than causal effects. Future research with access to firm-specific risk measures, temporal variation in policy stringency, or quasi-experimental settings would be valuable for establishing causality.

7. Conclusion

This study examines greenwashing in biodiversity reporting among Fortune Global 500 firms over 2017–2023. We quantify two distinct greenwashing practices leveraging state-of-the-art NLP techniques and analyze how these relate to different sources of legitimacy pressure. Our correlated random effects estimates show that cheap talk is positively associated with exposure to biodiversity-related financial risks, while positive talk is associated with the biodiversity policy stringency of firms' home countries. Both practices intensified around the Kunming–Montreal process, with cheap talk increasing earlier and positive talk strengthening more clearly following the adoption of the Global Biodiversity Framework, suggesting that firms strategically ad-

justed their disclosure practices as biodiversity-related institutional pressures escalated.

These findings suggest that biodiversity reporting may provide misleading signals of firms' risk management practices, raising concerns about information quality in financial markets. For investors, this underscores the critical need for careful scrutiny of biodiversity-related disclosures. For regulators and standard-setters, it highlights the urgent need to tighten reporting requirements and implement mechanisms to limit symbolic disclosures. For financial markets, our findings imply that greenwashing may already be distorting the assessment of biodiversity-related financial risks, potentially leading to their underpricing and the misallocation of capital. Such distortions can lead to abrupt valuation adjustments when risks crystallize, thereby amplifying transition shocks.

Beyond our empirical results, this study highlights the potential of NLP methods to enhance transparency within the highly complex informational environment of sustainability reporting. By transforming unstructured narratives into comparable indicators at scale, these technologies can enable investors, regulators, and scholars to more effectively distinguish substantive from symbolic disclosure. In doing so, they strengthen corporate accountability and help mitigate the information asymmetries that can distort the pricing of sustainability-related financial risks.

This study has three important limitations. First, it only examines two forms of greenwashing—cheap talk and positive talk—chosen for their suitability for detection through NLP. Other dimensions, such as omission of information, inaccurate reporting, or failure to act on commitments, were beyond the scope of this analysis. Second, the time-invariant nature of our key explanatory variables limited our ability to employ identification strategies that would support stronger causal inferences, meaning our findings should be interpreted as associations rather than causal relationships. Third, although our NLP models achieved a very high precision, classification inevitably involves some measurement error, which may affect the accuracy of our indicators.

References

[1] Adler, R., Mansi, M., Pandey, R., 2018. Biodiversity and threatened species reporting by the top Fortune Global companies. Accounting, Auditing & Accountability Journal 31, 787–825. URL:

https://www.emerald.com/insight/content/doi/10.1108/AAAJ-03-2016-2490/full/htm
doi:10.1108/AAAJ-03-2016-2490.

[2] Ali, R., García-Sánchez, I., Aibar-Guzmán, B., Rehman, R.U., 2024. Is biodiversity disclosure emerging as a key topic on the agenda of institutional investors? Business Strategy and the Environment 33, 2116–2142. URL: https://onlinelibrary.wiley.com/doi/10.1002/bse.3587, doi:10.1002/bse.3587.

[3] Ashforth, B.E., Gibbs, B.W., 1990. The Double-Edge of Organizational Legitimation. Organization Science 1, 177–194. URL: https://pubsonline.informs.org/doi/10.1287/orsc.1.2.177, doi:10.1287/orsc.1.2.177.

[4] Bafera, J., Kleinert, S., 2023. Signaling Theory in Entrepreneurship Research: A Systematic Review and Research Agenda. Entrepreneurship Theory and Practice 47, 2419–2464. URL: https://journals.sagepub.com/doi/10.1177/10422587221138489, doi:10.1177/10422587221138489.

[5] Bingler, J.A., Kraus, M., Leippold, M., Webersinke, N., 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. Journal of Banking & Finance 164, 107191. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378426624001080, doi:10.1016/j.jbankfin.2024.107191.

[6] Block, S., Emerson, J., Etsy, D., Sherbinin, A., Wendling, Z., 2024. Environmental Performance Index 2024. New Haven, CT.

[7] Boiral, O., 2016. Accounting for the Unaccountable: Biodiversity Reporting and Impression Management. Journal of Business Ethics 135, 751–768. URL: http://link.springer.com/10.1007/s10551-014-2497-9, doi:10.1007/s10551-014-2497-9.

[8] Christmann, P., Taylor, G., 2006. Firm self-regulation through international certifiable standards: determinants of symbolic versus substantive implementation. Journal of International Business Studies 37, 863–878. URL:

http://link.springer.com/10.1057/palgrave.jibs.8400231, doi:10.1057/palgrave.jibs.8400231.

[9] Connelly, B.L., Certo, S.T., Ireland, R.D., Reutzel, C.R., 2011. Signaling Theory: A Review and Assessment. Journal of Management 37, 39–67. URL: https://journals.sagepub.com/doi/10.1177/0149206310388419, doi:10.1177/0149206310388419.

[10] Cosma, S., Rimo, G., Cosma, S., 2023. Conservation finance: What are we not doing? A review and research agenda. Journal of Environmental Management 336, 117649. URL: https://linkinghub.elsevier.com/retrieve/pii/S0301479723004371, doi:10.1016/j.jenvman.2023.117649.

[11] Cowie, R.H., Bouchet, P., Fontaine, B., 2022. The Sixth Mass Extinction: fact, fiction or speculation? Biological Reviews 97, 640–663. URL: https://onlinelibrary.wiley.com/doi/10.1111/brv.12816, doi:10.1111/brv.12816.

[12] Dasgupta, P., 2021. The economics of biodiversity: the Dasgupta review: abridged version. Updated: 2 february 2021 ed., HM Treasury, London.

[13] De Grefte, J., De Bruin, B., 2025. Sustainable finance, biodiversity, and greenwashing: how contested values, metrics, and causation facilitate information distortion, information omission, and information pollution. Current Opinion in Environmental Sustainability 74, 101522. URL: https://linkinghub.elsevier.com/retrieve/pii/S1877343525000156, doi:10.1016/j.cosust.2025.101522.

[14] El Ghoul, S., Guedhami, O., Kim, H., Park, K., 2018. Corporate Environmental Responsibility and the Cost of Capital: International Evidence. Journal of Business Ethics 149, 335–361. URL: http://link.springer.com/10.1007/s10551-015-3005-6, doi:10.1007/s10551-015-3005-6.

[15] Elliot, V., Jonäll, K., Paananen, M., Bebbington, J., Michelon, G., 2024. Biodiversity reporting: standardization, materiality, and assurance. Current Opinion in Environmental Sustainability 68, 101435. URL:

https://linkinghub.elsevier.com/retrieve/pii/S1877343524000228, doi:10.1016/j.cosust.2024.101435.

[16] EPI, Y., 2022. URL: https://epi.yale.edu/epi-results/2022/component/bdh.

[17] Garel, A., Romec, A., Sautner, Z., Wagner, A.F., 2024. Do investors care about biodiversity? Review of Finance 28, 1151–1186. URL: https://academic.oup.com/rof/article/28/4/1151/7645412, doi:10.1093/rof/rfae010.

[18] Giglio, S., Kuchler, T., Stroebel, J., Zeng, X., 2023. Biodiversity Risk. Technical Report w31137. National Bureau of Economic Research. Cambridge, MA. URL: http://www.nber.org/papers/w31137.pdf, doi:10.3386/w31137.

[19] Gjerde, S., Sautner, Z., Wagner, A.F., Wegerich, A., 2025. Corporate nature risk perceptions. Review of Finance URL: https://ssrn.com/abstract=5224376, doi:10.2139/ssrn.5224376. forthcoming.

[20] Guer, Y., Mueller, L., Schiereck, D., 2024. Biodiversity: A Bibliometric Analysis of Accounting, Economics, and Finance Journals. Journal of Alternative Finance 1, 194–210. URL: https://journals.sagepub.com/doi/10.1177/27533743241236802, doi:10.1177/27533743241236802.

[21] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 8342–8360. URL: https://www.aclweb.org/anthology/2020.acl-main.740, doi:10.18653/v1/2020.acl-main.740.

[22] Hassan, A.M., Roberts, L., Atkins, J., 2020. Exploring factors relating to extinction disclosures: What motivates companies to report on biodiversity and species protection? Business Strategy and the Environment 29, 1419–1436. URL: https://onlinelibrary.wiley.com/doi/10.1002/bse.2442, doi:10.1002/bse.2442.

[23] Hudson, R., 2024. Biodiversity and risk in the financial sector. Journal of Sustainable Finance and Accounting 2, 100009. URL: https://linkinghub.elsevier.com/retrieve/pii/S2950370124000099, doi:10.1016/j.josfa.2024.100009.

[24] Hutchinson, M.C., Lucey, B., 2024. A bibliometric and systemic literature review of biodiversity finance. Finance Research Letters 64, 105377. URL: https://linkinghub.elsevier.com/retrieve/pii/S1544612324004070, doi:10.1016/j.frl.2024.105377.

[25] Jonäll, K., Baeckström, Y., Elliot, V., Arvidsson, S., 2025. The biodiversity–finance nexus: a future research agenda. Current Opinion in Environmental Sustainability 72, 101504. URL: https://linkinghub.elsevier.com/retrieve/pii/S1877343524000915, doi:10.1016/j.cosust.2024.101504.

[26] Karolyi, G.A., Tobin-de La Puente, J., 2023. Biodiversity finance: A call for research into financing nature. Financial Management 52, 231–251. URL: https://onlinelibrary.wiley.com/doi/10.1111/fima.12417, doi:10.1111/fima.12417.

[27] Kennedy, S., Fuchs, M., Van Ingen, W., Schoenmaker, D., 2023. A resilience approach to corporate biodiversity impact measurement. Business Strategy and the Environment 32, 2567–2582. URL: https://onlinelibrary.wiley.com/doi/10.1002/bse.3140, doi:10.1002/bse.3140.

[28] Kirtac, K., Germano, G., 2025. Large language models in finance: what is financial sentiment? arXiv preprint arXiv:2503.03612. URL: https://arxiv.org/abs/2503.03612, doi:10.48550/arXiv.2503.03612.

[29] Mair, L., Elnahass, M., Xiang, E., Hawkins, F., Siikamaki, J., Hillis, L., Barrie, S., McGowan, P.J.K., 2024. Corporate disclosures need a biodiversity outcome focus and regulatory backing to deliver global conservation goals. Conservation Letters 17, e13024. URL: https://conbio.onlinelibrary.wiley.com/doi/10.1111/conl.13024, doi:10.1111/conl.13024.

[30] Marquis, C., Toffel, M.W., Zhou, Y., 2016. Scrutiny, Norms, and Selective Disclosure: A Global Study of Greenwashing. Organization Science 27, 483–504. URL: https://pubsonline.informs.org/doi/10.1287/orsc.2015.1039, doi:10.1287/orsc.2015.1039.

[31] Mundlak, Y., 1978. On the Pooling of Time Series and Cross Section Data. Econometrica 46, 69. URL: https://www.jstor.org/stable/1913646?origin=crossref, doi:10.2307/1913646.

[32] Payne, J.L., Bush, A.M., Heim, N.A., Knope, M.L., McCauley, D.J., 2016. Ecological selectivity of the emerging mass extinction in the oceans. Science 353, 1284–1286. URL: https://www.science.org/doi/10.1126/science.aaf2416, doi:10.1126/science.aaf2416.

[33] Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. URL: https://arxiv.org/abs/1910.01108, doi:10.48550/ARXIV.1910.01108. version Number: 4.

[34] Schimanski, T., Colesanti Senni, C., Gostlow, G., Ni, J., Yu, T., Leippold, M., 2023. Exploring nature: Datasets and models for analyzing nature-related disclosures. arXiv preprint arXiv:2312.17337 URL: https://arxiv.org/abs/2312.17337, doi:10.48550/arXiv.2312.17337.

[35] Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., Leippold, M., 2024. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. Finance Research Letters 61, 104979. URL: https://linkinghub.elsevier.com/retrieve/pii/S1544612324000096, doi:10.1016/j.frl.2024.104979.

[36] Shakourloo, A., Azimli, A., 2025. The role of greenwashing scandals on investors' herding behavior. Journal of Economics and Finance URL: https://link.springer.com/10.1007/s12197-025-09725-x, doi:10.1007/s12197-025-09725-x.

[37] Suchman, M.C., 1995. Managing Legitimacy: Strategic and Institutional Approaches. The Academy of Management Review 20, 571. URL: http://www.jstor.org/stable/258788?origin=crossref, doi:10.2307/258788.

[38] Svartzman, R., Espagne, E., Julien, G., Paul, H.L., Mathilde, S., Allen, T., Berger, J., Calas, J., Godin, A., Vallier, A., 2021. A 'Silent Spring' for the Financial System? Exploring Biodiversity-Related Financial Risks in France. SSRN Electronic Journal URL: https://www.ssrn.com/abstract=4028442, doi:10.2139/ssrn.4028442.

[39] Vamvourellis, D., Mehta, D., 2025. Reasoning or overthinking: Evaluating large language models on financial sentiment analysis. arXiv preprint arXiv:2506.04574. URL: https://arxiv.org/abs/2506.04574, doi:10.48550/arXiv.2506.04574.

[40] Walker, K., Wan, F., 2012. The Harm of Symbolic Actions and Green-Washing: Corporate Actions and Communications on Environmental Performance and Their Financial Implications. Journal of Business Ethics 109, 227–242. URL: http://link.springer.com/10.1007/s10551-011-1122-4, doi:10.1007/s10551-011-1122-4.

[41] Webersinke, N., Kraus, M., Bingler, J.A., Leippold, M., 2021. ClimateBert: A Pretrained Language Model for Climate-Related Text. URL: https://arxiv.org/abs/2110.12010, doi:10.48550/ARXIV.2110.12010. version Number: 3.

[42] Wooldridge, J.M., 2019. Correlated random effects models with unbalanced panels. Journal of Econometrics 211, 137–150. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407618302392, doi:10.1016/j.jeconom.2018.12.010.

[43] Yadav, P.L., Han, S.H., Rho, J.J., 2016. Impact of Environmental Performance on Firm Value for Sustainable Investment: Evidence from Large US Firms. Business Strategy and the Environment 25, 402–420. URL: https://onlinelibrary.wiley.com/doi/10.1002/bse.1883, doi:10.1002/bse.1883.

[44] Zu Ermgassen, S.O., Howard, M., Bennun, L., Addison, P.F., Bull, J.W., Loveridge, R., Pollard, E., Starkey, M., 2022. Are corporate biodiversity commitments consistent with delivering 'nature-positive' outcomes? A review of 'nature-positive' definitions, company progress and challenges. Journal of Cleaner Production 379, 134798. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959652622043700, doi:10.1016/j.jclepro.2022.134798.

Appendix A. Biodiversity classifier evaluation

To ensure that the biodiversity sentence classifier used in the first stage of our NLP pipeline performs reliably on our sample, we conducted an independent out-of-sample validation. Although the ESGBERT biodiversity model reports strong performance in its original evaluation, it was necessary to verify that these performance levels hold within the specific context of Fortune Global 500 sustainability reporting.

The authors of the ESGBERT-biodiversity model describe employing corporate sustainability reports in its training (along with annual reports, and corporate news). However, they do not specify which firm-year reports were included in the training. To ensure a rigorous out-of-sample evaluation, we therefore restrict our validation set to FY 2023 sustainability reports, published in 2024, after the model's release date. This temporal separation ensures that our evaluation reflects genuine predictive performance on previously unseen documents, rather than any potential overlap with the training data.

We implemented a multi-stage, stratified sampling strategy to construct a validation dataset that captures the full range of contexts in which biodiversity reporting occurs, and that includes the varied edge cases the classifier must distinguish. First, we selected 100 sustainability reports using proportional stratified sampling across industry–country combinations to ensure broad coverage of corporate reporting contexts. Because biodiversity-related sentences are relatively sparse in these documents, a purely random sentence sample would yield very few positive cases. To address this, we extracted sentences using the comprehensive keyword dictionary developed by Schimanski et al. (2024), which spans four thematic categories: water-related terms, forest and agriculture- related terms, biodiversity-related terms, and general environment-related terms. This procedure ensured that the validation set included both clearly biodiversity-related content and the more ambiguous edge cases that the classifier must distinguish in practice.

Following initial extraction, we applied a two-level balancing procedure: first maintaining proportional representation across industry-country strata, then equalizing the distribution across the four biodiversity categories (targeting 25% per category) while preserving the underlying stratification structure. This dual-balancing approach yielded a dataset that is both representative of the corporate landscape and balanced in terms of thematic biodiversity content, capturing the full spectrum from explicit biodiversity mentions to

more ambiguous environmental references.

Two researchers independently labeled each sentence as biodiversity-related or not. Inter-rater agreement was high (Cohen's k = 0.938). Disagreements (n = 16, 1.6%) were resolved through discussion and, when necessary, consultation of the original report context.

All sentences were later classified by the ESGBERT-Biodiversity model, and performance metrics were subsequently calculated. Both the original and validation metrics are presented in Table A.1.

Table A.1 – Biodiversity Sentence Classifier Performance: Original vs. Validation Metrics

Metric	Original (Schimanski et al., 2023)	Validation
F1-score (sd)	0.9276 (0.0191)	0.9453 (0.0131)
Accuracy (sd)	0.9582 (0.0105)	0.9830 (0.0040)
Precision (sd)	0.9285 (0.0176)	0.9671 (0.0136)
Recall (sd)	0.9278 (0.0267)	0.9245 (0.0207)

The confusion matrix presented in Table A.2 demonstrates high classifier reliability. Of 159 manually labeled biodiversity sentences, the model correctly identified 147 (92.5%), with only 12 false negatives. Critically, false positives were minimal: only 5 of 841 non-biodiversity sentences were misclassified, yielding 96.7% precision. This high precision ensures that generic environmental content does not contaminate the biodiversity-specific sample and distort downstream greenwashing measures. With only 5 false positives (0.5% of the validation sample), we can be confident that the Cheap Talk and Positive Talk Indices are computed on genuinely biodiversity-related content rather than general environmental discourse.

Table A.2 – Confusion Matrix – Biodiversity Classifier Validation

	Predicted Biodiversity	Predicted Not Biodiversity
True Biodiversity	147	12
True Not Biodiversity	5	836

To evaluate the model's reliability when applied to generic sustainability reporting language, we conducted an additional test using 500 randomly

sampled sentences. Because this corpus is overwhelmingly composed of non-biodiversity-related content, the exercise specifically assesses the classifier's ability to avoid false positives in generic corporate reporting text. The model achieved perfect performance in this setting, correctly classifying every sentence in the sample. This demonstrates that the classifier does not mistakenly flag generic corporate language as biodiversity-related.

In our study, biodiversity relevance is determined at the paragraph level using a ≥ 1 sentence aggregation rule: any paragraph containing at least one biodiversity-related sentence enters the analysis. To assess how sentence-level classification errors propagate under this aggregation, we examine paragraph-level behavior within the out-of-sample validation set.

At the sentence level, the classifier correctly identifies 147 of 159 true biodiversity sentences, with 12 false negatives. When predictions are aggregated at the paragraph level, 4 of these sentence-level false negatives are recovered, as they occur in paragraphs that also contain at least one biodiversity sentence correctly classified by the model. As a result, paragraph-level recall exceeds sentence-level recall, indicating that the aggregation rule mitigates sentence-level false negatives.

A similar pattern is observed for false positives. Of the five sentences misclassified as biodiversity-related at the sentence level, two occur in paragraphs that are genuinely biodiversity-related and therefore do not translate into paragraph-level false positives.

At both margins, recovered false negatives and false positives correspond to sentence-level edge cases that are difficult to classify in isolation. In these cases, individual sentences are borderline when considered on their own, but appear within paragraphs that contain other clearly biodiversity-related sentences correctly identified by the model.

Overall, these results show that the $\geq$1 sentence aggregation rule dampens sentence-level errors in both directions, improving paragraph-level recall while maintaining very high precision and ensuring that biodiversity-relevant paragraphs are retained in the analytical sample.

Because validation is conducted at the sentence level, paragraph-level true negatives are not a well-defined object of analysis. A sentence that is correctly classified as non-biodiversity does not imply that the paragraph to which it belongs is non-biodiversity, since the same paragraph may contain other biodiversity-related sentences that are correctly identified by the model but not part of the validation set. As a result, paragraph-level true negatives cannot be reliably enumerated without exhaustively annotating all sentences

within each paragraph. For this reason, we do not report a full paragraph-level confusion matrix. Instead, we concentrate on how sentence-level false negatives and false positives propagate under the ≥ 1 sentence aggregation rule.

Appendix B. Data annotation

1. Commitment Classification

We adopted the commitment definition employed by Bingler et al. (2024), which is both forward- and backward-looking. Under this definition, commitments include both completed actions and future targets. Additionally, commitments cover not only activities that organizations actively pursue but also those they deliberately avoid (e.g., ceasing to purchase goods from suppliers connected to deforestation).

A paragraph is classified as a commitment if it contains information related to:

- Actions undertaken and/or targets achieved related to biodiversity protection or impact mitigation.
- Planned actions, targets set, and/or stated intentions to act related to biodiversity protection or impact mitigation.

A paragraph is classified as non-commitment if it does not contain information about biodiversity-related actions, targets, or intentions.

Commitment example: *We collaborated with the Consumer Goods Forum to develop a methodology for assessing anti-deforestation policies among traders in its supply chains. In 2021, 31 traders were evaluated, and ongoing efforts are focused on strengthening the reliability of the assessment criteria so that trader progress can be measured in 2022.*

Non-commitment example: *Land degradation encompasses the range of processes that lead to declines in biodiversity, ecosystem functions, or their benefits to people, and includes the deterioration of all terrestrial ecosystems.*

2. Specificity Classification

Following Bingler et al. (2024) and Subramanian et al. (2019), commitment paragraphs are classified as specific if they meet one or more of the following conditions:

- Define verifiable goals in the past, present, or future.

- Detail specific actions for achieving a tangible objective.
- Detail how goals were/will be tracked.

Conversely, commitment paragraphs are classified as non-specific if they:

- Define intangible, non-verifiable goals.
- Define broad goals without any elaboration on how to achieve or track them.
- Contain general statements that apply to all situations or the entire sector.
- Discuss public policies or regulations without being specific about what the company has implemented or plans to implement.

Specific commitment example:*Our commitment is particularly relevant for soybeans sourced from high-risk deforestation areas in South America, including the Brazilian Cerrado and the Argentinian Chaco. By 2020, all soy volumes directly sourced from farms in these regions were fully traceable.*

Non-specific commitment example:*As part of its commitment to avoiding potential environmental harm, the Company undertakes all reasonable efforts to safeguard, conserve, and restore natural resources.*

3. Sentiment Classification

Following Bingler et al. (2024), paragraphs are classified as having:

- Negative sentiment - If they address actual or potential risks, losses, or adverse developments that impact a firm; describe a firm's actual or potential negative impacts on biodiversity; or employ negative adjectives to describe biodiversity-related developments.
- Positive sentiment - If they describe business opportunities from biodiversity, address positive impacts, address potential reductions of negative impacts, or employ positive adjectives.
- Neutral sentiment - If they state biodiversity-related facts without framing them positively or negatively.

Positive sentiment example: *We consider that the growing attention to biodiversity and the shift towards a recycling-oriented society create new business opportunities. We will continue to support the pursuit of a sustainable society through our business activities.*

Neutral sentiment example: *The TNFD is a risk management and disclosure framework that allows companies to identify, manage, and disclose their impacts on natural capital, including environmental degradation and biodiversity loss.*

Negative sentiment example: *Nature and biodiversity loss represent a risk for our group, as they can undermine the availability and quality of key inputs and heighten exposure to climate-related impacts. These factors may, in turn, result in supply disruptions, regulatory intervention, and reputational impacts.*

Appendix C. Classifier Cross-Validation Performance

This section reports classification performance from 5-fold cross-validation for all three classifiers. For each task, we present per-class precision, recall, and F1 scores averaged across folds, with standard deviations reported in parentheses. In addition, pooled out-of-sample confusion matrices constructed from cross-validation predictions are provided to illustrate systematic patterns of misclassification.

Appendix C.1. Per-class performance metrics

Table A.3 – Per-class performance metrics – commitment classifier

	Precision (mean ± sd)	**Recall (mean ± sd)**	**F1 (mean ± sd)**
Non-commitment	0.922 ± 0.050	0.862 ± 0.036	0.890 ± 0.024
Commitment	0.934 ± 0.015	0.962 ± 0.027	0.947 ± 0.012

Table A.4 – Per-class performance metrics – specificity classifier

	Precision (mean ± sd)	**Recall (mean ± sd)**	**F1 (mean ± sd)**
Non-specific	0.884 ± 0.029	0.849 ± 0.009	0.866 ± 0.016
Specific	0.824 ± 0.011	0.864 ± 0.038	0.843 ± 0.023

Table A.5 – Per-class performance metrics – sentiment classifier

	Precision (mean ± sd)	Recall (mean ± sd)	F1 (mean ± sd)
Positive	0.931 ± 0.013	0.963 ± 0.021	0.947 ± 0.013
Neutral	0.841 ± 0.070	0.735 ± 0.022	0.783 ± 0.030
Negative	0.829 ± 0.068	0.833 ± 0.031	0.829 ± 0.035

Appendix C.2. Confusion matrices

Table A.6 – Pooled out-of-sample confusion matrix – commitment classifier

	Predicted non-commitment	Predicted commitment
True non-commitment	574	91
True commitment	51	1284

Table A.7 – Pooled out-of-sample confusion matrix – specificity classifier

	Predicted non-specific	Predicted specific
True non-specific	623	111
True specific	82	521

Table A.8 – Pooled out-of-sample confusion matrix – sentiment classifier

	Predicted positive	Predicted neutral	Predicted negative
True positive	1341	39	12
True neutral	83	300	23
True negative	15	18	169

Appendix C.3. Classifier performance summary and error structure

Overall, the three classifiers exhibit strong and stable performance under 5-fold cross-validation, with high precision and recall for the primary classes of interest and low dispersion across folds. The magnitude and distribution of the reported performance metrics are consistent with those documented for the original classifier architectures developed by Bingler et al. (2024), which

form the basis of the fine-tuned models employed in this study. Performance is consistently strongest for the commitment and positive sentiment classes, reflecting the clearer semantic structure of explicit commitments and evaluative language in corporate disclosures. The commitment classifier achieves balanced performance across both classes (F1 = 0.890 and 0.947), while the sentiment classifier demonstrates particularly strong discrimination of positive content (F1 = 0.947).

In contrast, neutral and non-specific classes display comparatively lower performance, consistent with their inherently more ambiguous linguistic boundaries. Neutral sentiment, in particular, occupies a conceptually intermediate position between positive and negative evaluations, often characterised by factual statements, hedged language, or mixed signals that defy clear classification. This conceptual ambiguity is reflected empirically: the pooled confusion matrix shows that neutral misclassifications are distributed across both positive (84 instances) and negative (24 instances) categories rather than systematically concentrated in one direction. Such bidirectional confusion is characteristic of boundary cases rather than systematic bias and is well documented in the finance-related NLP literature, where even state-of-the-art models struggle to distinguish neutral from weakly toned sentiment [39, 28]).

Finally, the confusion matrices indicate that classification errors arise predominantly between adjacent or semantically proximate categories rather than from fundamental misidentification of construct-relevant content. In the sentiment classifier, confusion between positive and negative poles is minimal, with most errors involving the neutral category. Similarly, in the binary commitment task, cross-class errors are limited relative to class prevalence, with the vast majority of commitment statements correctly identified and only a small share of non-commitments misclassified as commitments. This error structure suggests that the classifiers reliably capture the core theoretical constructs while exhibiting predictable uncertainty at category boundaries, a pattern consistent with known challenges in human annotation. Taken together, these results indicate that the classifiers provide valid and reliable measurements of the intended constructs, with error patterns that are both interpretable and theoretically consistent with the semantic structure of corporate climate disclosure.

Appendix D. Classifier Audit and Qualitative Error Analysis

To validate the performance of our fine-tuned classification models we conducted a blinded human audit of model predictions. This exercise complements the cross-validation metrics by providing an independent assessment of model accuracy on out-of-sample data drawn from the actual dataset used in our analysis. We constructed a random sample of 600 biodiversity-related paragraphs from sustainability reports published between 2017 and 2023, stratified by year, country, and industry. Two researchers independently coded all 600 paragraphs following the annotation guidelines described in Appendix B. Cohen's kappa values indicate substantial agreement across all three classification tasks (Commitment:0.89, Specificity:0.79, Sentiment: 0.87), providing confidence in the reliability of human judgment and the clarity of the annotation guidelines.

Tables A.9–A.11 present confusion matrices comparing model predictions against ground truth labels. The audit confirms solid classifier performance across all tasks, with accuracies of 93.8% (commitment), 89.0% (specificity), and 95.5% (sentiment). Importantly, these results validate the construction of our key indicators: the specificity classifier exhibits a conservative bias, meaning the CTI likely underestimates rather than overestimates vague commitment language. The sentiment classifier demonstrates minimal confusion between positive and negative poles (only 3 of 600 paragraphs), with most errors occurring in the neutral category. This provides strong confidence in the validity of the PTI, which relies exclusively on the positive-to-negative ratio.

Table A.9 – Confusion matrix – Commitment classifier Audit

	Predicted non-commitment	Predicted commitment
True non-commitment	86	22
True commitment	15	477

Table A.10 – Confusion matrix – Specificity Classifier Audit

	Predicted non-specific	Predicted specific
True non-specific	249	43
True specific	11	189

Table A.11 – Confusion Matrix – Sentiment Classifier Audit

	Predicted positive	Predicted neutral	Predicted negative
True positive	486	12	1
True neutral	9	46	0
True negative	1	4	41

To further understand the nature of misclassifications, we conducted a qualitative review of all audit errors. This analysis indicates that misclassifications predominantly arise in genuinely ambiguous cases, where the distinction between categories is subtle and context-dependent.

In the commitment classifier, false positives typically consist of paragraphs that signal intent or concern without clearly articulating a concrete action, target, or intention, as illustrated by the two paraphrased examples below.

- *We fly to some of the most beautiful places on the planet. Hence, we care deeply about ensuring that the natural habitats we rely on remain healthy for the long haul.*
- *We acknowledge the need to look beyond responsible sourcing in order to support biodiversity, ecosystems, and the communities that depend on natural landscapes.*

These paragraphs may reasonably be interpreted as signalling intent or concern, even though they do not directly articulate a commitment under our annotation guidelines. Notably, almost no false positives consist of purely descriptive policy, certification, or regulatory background paragraphs, indicating that the model does not systematically misclassify generic contextual information as commitments.

Similarly, false negatives in the commitment classifier primarily involve weakly articulated or indirectly framed commitments, as illustrated by the following paraphrased examples:

- *We recognise growing societal concern about marine pollution and other environmental issues and are therefore committed to reducing the harmful impacts brought by human activities.*

- *As a company with an agricultural supply chain, contributing to protecting nature and the ecosystems where we operate is essential.*

In the specificity classifier, misclassifications again concentrate in borderline cases. We identify two main types of false positives. The first consists of broad and ambitious commitments that include concrete targets or timelines but lack detail on implementation or monitoring, for example:

- *We unveiled a new goal for all of our direct and indirect supply chains to be free of native species deforestation in high-risk areas by December 31, 2021, as we continue on our pathway to become 100% deforestation-free.*

The second type comprises statements that provide rich descriptive detail about projects—such as locations, species, ecosystems, or partner organisations—without specifying concrete actions, outcomes, or metrics, for example:

- *We continued our Rio Negro Estuary Rescue Project in San Luis Potosí in 2020, where an abundance of ichthyofauna was identified.*

False negatives in the specificity classifier similarly reflect borderline cases in which concrete actions or timelines are mentioned, but the scope, intensity, or quality of implementation remains unclear, as in:

- *We are committed to implementing plans to manage pharmaceutical residues in the environment and their potential impact on ecosystems in all our production sites by 2023.*

Two additional types of false negatives merit discussion. First, absolute commitments to cease involvement in specific activities are sometimes classified as non-specific, as in:

- *Given the adverse implications for ecosystems, new business activities—including insurance underwriting, investment, and financing—will not be pursued in relation to oil sands mining projects.*

Because commitments framed in this way frequently involve broad ambitions (e.g. avoiding deforestation or illegal fishing) without implementation detail, the classifier treats them as non-specific. In some contexts, however, such statements may be substantively precise. These cases are rare and do not materially affect the Cheap Talk Index.

Second, highly specific initiatives that are difficult to quantify may also be classified as non-specific, for example:

- *We began collaborating with the indigenous communities of Sei Talao and Sei Kunyit to protect our conservation reserve in Sumatra, with a focus on cleaning the Sawit River basin for preserving its water quality.*

Finally, errors in the sentiment classifier are most commonly observed in paragraphs that can plausibly be read as either neutral description or positive framing, depending on interpretation. An illustrative example is:

- *Protecting the planet's biodiversity is a global imperative, with the focus now shifting from mitigating damage to achieving a positive impact.*

Overall, the qualitative error analysis indicates that misclassifications arise primarily in ambiguous cases near conceptual category boundaries rather than from systematic bias. Importantly, no error pattern suggests a directional distortion that would mechanically generate or exaggerate the study's main empirical results. If anything, the observed error structure implies conservative measurement of both the CTI and the PTI.

Appendix E. Classifier Calibration and Cross-strata Performance

This appendix reports additional diagnostic checks on the NLP classifiers used to construct the disclosure-based indicators. We assess probability calibration, using reliability diagrams and Brier scores, and performance stability across strata, reporting standard classification metrics by region and industry. These diagnostics document the reliability of predicted probabilities and the robustness of classification performance across the main dimensions along which the empirical analysis varies. All results are based on the manually annotated out-of-sample validation set.

Appendix E.1. Biodiversity sentence classifier

Figure E.6 reports the calibration curve for the biodiversity sentence classifier using quantile bins. The curve closely follows the 45-degree line across most of the probability range, indicating a very good alignment between predicted probabilities and observed frequencies. At low predicted probabilities, observed frequencies remain close to zero, suggesting limited false activation on generic environmental or corporate language. At higher predicted probabilities, observed frequencies increase smoothly and remain close to predicted values, with minimal deviations at the upper end of the distribution.

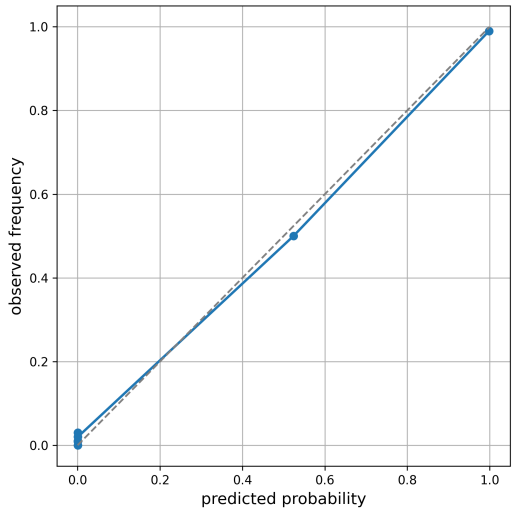


Figure E.6: Calibration curve for the biodiversity sentence classifier

Consistent with the visual evidence, the classifier achieves a very low Brier score (0.018), indicating small average squared deviations between predicted probabilities and realised outcomes. At the operational threshold of 0.5, overall performance is strong (precision = 0.96, recall = 0.92, F1 = 0.94). Performance is stable across industries despite substantial differences in baseline prevalence, with F1-scores ranging from 0.92 to 1 (Table A.12). The region-level results similarly show robust performance across geographies, with no indication of systematic degradation (Table A.13).

Table A.12. Performance of the biodiversity classifier by industry

Industry	Observations	Prevalence	Precision	Recall	F1-score
Consumer staples	148	0.115	0.909	0.930	0.919
Consumer discretionary	84	0.131	0.993	0.955	0.973
Energy	143	0.280	0.990	0.975	0.982
Financials	220	0.186	0.986	0.939	0.961
Industrials	187	0.118	1.000	1.000	1.000
Materials	65	0.154	0.982	0.900	0.936
Other	153	0.118	0.950	0.993	0.970

Table A.13. Performance of the biodiversity classifier by region

Region	Observations	Prevalence	Precision	Recall	F1-score
Asia	339	0.159	0.983	0.961	0.972
Europe	439	0.178	0.970	0.951	0.960
North America	155	0.110	1.000	1.000	1.000
Other	67	0.149	0.941	0.941	0.941

Appendix E.2. Commitment classifier

Figure E.7 presents the calibration curve for the commitment classifier using quantile bins. The curve closely tracks the diagonal across most of the probability distribution, particularly at higher predicted probabilities where the majority of observations are concentrated. This pattern indicates strong probability alignment in the region most relevant for classification decisions. At lower predicted probabilities, observed frequencies lie slightly above the diagonal, indicating mild underconfidence in this range. These bins represent a small portion of the probability mass and are unlikely to materially affect classification at the operational threshold of 0.5. The Brier score (0.055) confirms good probability calibration. Overall classification performance is very high (precision = 0.95, recall = 0.97, F1 = 0.96).

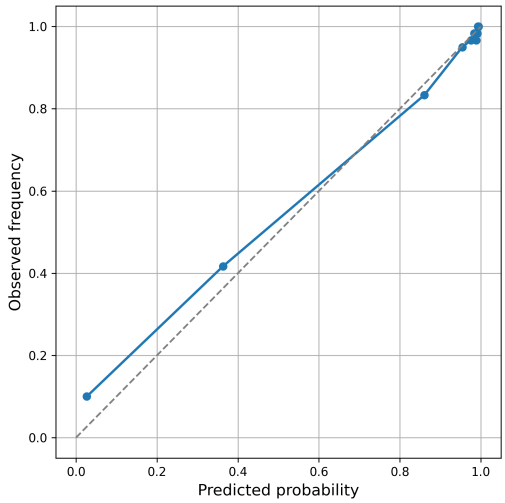


Figure E.7: Calibration curve for the commitment classifier

Performance remains consistently strong across regions and industries, with F1-scores exceeding 0.90 in all strata (Tables A.14-15). This stability indicates that commitment detection is robust across heterogeneous reporting contexts and provides a reliable input for downstream analysis.

Table A.14. Performance of the commitment classifier by industry

Industry	Observations	Prevalence	Precision	Recall	F1-score
Consumer staples	127	0.819	0.926	0.962	0.943
Consumer discretionary	49	0.755	0.923	0.973	0.947
Energy	161	0.826	0.970	0.962	0.966
Financials	64	0.781	0.980	1.000	0.990
Industrials	54	0.907	1.000	0.959	0.979
Materials	46	0.826	0.925	0.974	0.949
Other	99	0.818	0.963	0.975	0.969

Table A.15. Performance of the commitment classifier by region

Region	Observations	Prevalence	Precision	Recall	F1-score
Asia	199	0.809	0.945	0.969	0.957
Europe	226	0.792	0.972	0.961	0.966
North America	146	0.904	0.970	0.977	0.974
Other	29	0.690	0.833	1.000	0.909

Appendix E.3. Specificity classifier

Figure E.8 reports the calibration curve for the specificity classifier using quantile bins. Relative to the other classifiers, the curve exhibits systematic overconfidence at intermediate predicted probabilities, where predicted values exceed observed frequencies. This region corresponds to semantically ambiguous, borderline cases in which distinctions between specific and non-specific commitments are inherently difficult, as documented in the qualitative analysis presented in the previous appendix. At higher predicted probabilities—where the bulk of classification mass is concentrated—the curve closely follows the 45-degree line, indicating good calibration for high-confidence predictions.

Importantly, this calibration profile implies a conservative bias for the Cheap Talk Index. Overidentification of specificity mechanically reduces the numerator of the index (number of non-specific commitment paragraphs), biasing estimates of cheap talk downward rather than inflating them.

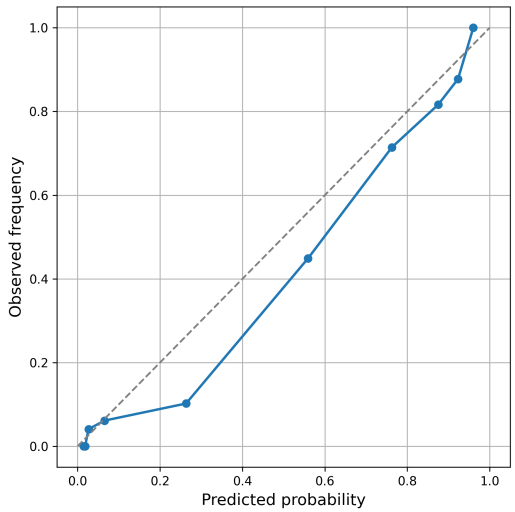


Figure E.8: Calibration curve for the specificity classifier

The pattern observed in the diagram is consistent with the higher Brier score for specificity (0.091). The classifier exhibits high recall (0.95) and good precision (0.81), yielding an F1-score of 0.88. Performance remains stable across regions and industries, with F1-scores ranging between 0.79 and 0.94 and no evidence of systematic failure in any particular stratum (Tables A.16-17).

Table A.16. Performance of the specificity classifier by industry

Industry	Observations	Prevalence	Precision	Recall	F1-score
Consumer staples	104	0.490	0.836	1.000	0.911
Consumer discretionary	37	0.324	0.750	1.000	0.857
Energy	132	0.417	0.812	0.945	0.874
Financials	50	0.460	0.917	0.957	0.936
Industrials	49	0.367	0.727	0.889	0.800
Materials	38	0.368	0.786	0.786	0.786
Other	82	0.329	0.806	0.926	0.862

Table A.17. Performance of the specificity classifier by region

Region	Observations	Prevalence	Precision	Recall	F1-score
Asia	160	0.350	0.820	0.893	0.855
Europe	179	0.369	0.797	0.955	0.869
North America	132	0.485	0.810	1.000	0.895
Other	21	0.667	0.923	0.857	0.889

Appendix E.4. Sentiment classifier

Figures E.9-11 report one-vs-rest calibration curves for the negative, positive, and neutral sentiment classes. Interpretation focuses on the negative and positive classes, which directly enter the construction of the Positive Talk Index, while neutral sentiment is reported for completeness.

The calibration curve for negative sentiment (Figure E.9) closely follows the 45-degree line across the probability range, indicating a very good alignment between predicted probabilities and observed frequencies. Deviations at very low predicted probabilities are minimal, and mid-range predictions closely match realised outcomes. This pattern is consistent with the low Brier score for the negative class (0.010) and supports reliable identification of negative biodiversity-related disclosures.

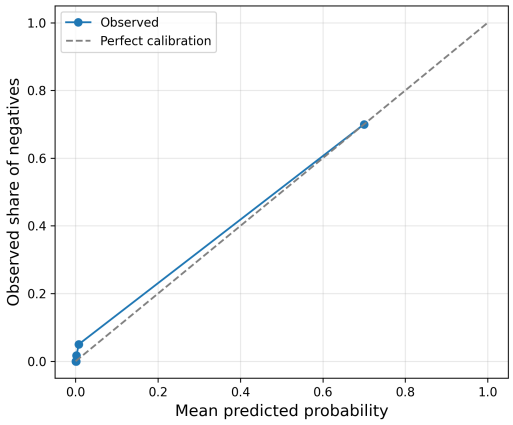


Figure E.9: Calibration curve for the sentiment classifier (negative)

For positive sentiment (Figure E.10), calibration is strong at higher predicted probabilities, where most observations are concentrated. At intermediate probability levels, observed frequencies lie slightly above the diagonal,

indicating mild underconfidence. Calibration converges toward the diagonal at higher confidence levels, which are most relevant for classification decisions. The Brier score for the positive class is correspondingly low (0.033).

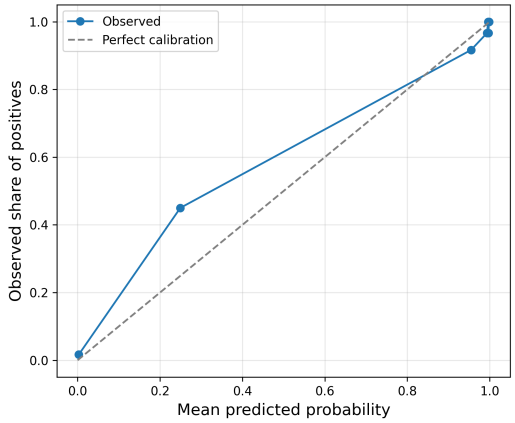


Figure E.10: Calibration curve for the sentiment classifier (positive)

The neutral sentiment curve (Figure E.11) exhibits deviations from perfect calibration, particularly at higher predicted probabilities, reflecting the well-documented difficulty of correctly distinguishing between neutral and borderline positive and negative cases in corporate disclosures. The Brier score for the neutral class is also low (0.037).

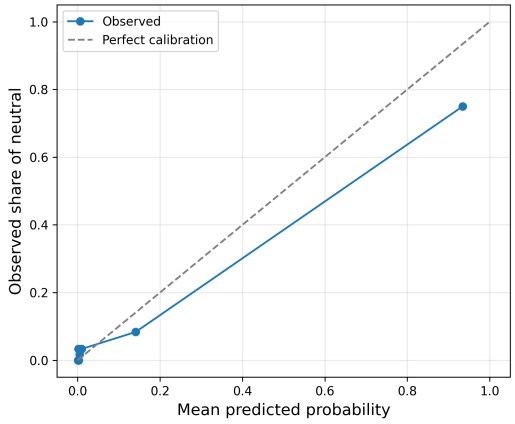


Figure E.11: Calibration curve for the sentiment classifier (neutral)

Overall performance of the sentiment classifier is strong, with high class-specific performance for the positive (precision = 0.98 , recall = 0.97, F1 = 0.98) and negative sentiment classes (precision = 0.98, recall = 0.89, F1 = 0.93). Stratified results by region and industry, reported as macro-averaged precision, recall, and F1-scores across the three sentiment classes, indicate stable performance (Tables A.18-19).

Table A.18. Performance of the sentiment classifier by industry

Industry	Observations	Precision	Recall	F1-score
Consumer discretionary	49	0.936	0.925	0.924
Consumer staples	127	0.856	0.801	0.826
Energy	161	0.920	0.920	0.920
Financials	64	0.833	0.934	0.854
Industrials	54	0.833	0.910	0.838
Materials	46	1.000	1.000	1.000
Other	99	0.952	0.996	0.972

Notes: Reported values are macro-averaged across the positive, neutral, and negative sentiment classes.

Table A.19. Performance of the sentiment classifier by region

Region	Observations	Precision	Recall	F1-score
Asia	199	0.913	0.948	0.929
Europe	226	0.879	0.892	0.883
North America	146	0.993	0.859	0.918
Other	29	0.873	0.806	0.827

Notes: Reported values are macro-averaged across the positive, neutral, and negative sentiment classes.

Appendix E.5. Summary

Across all classifiers, calibration diagnostics and per-stratum performance metrics indicate stable behavior across regions and industries, with no evidence of systematic sectoral or geographic degradation. The modest calibration deviation observed for the specificity classifier operates in a conservative direction with respect to the construction of the Cheap Talk Index. These results indicate that the Cheap Talk and Positive Talk indices are grounded

in reliable, well-calibrated predictions and are unlikely to be mechanically driven by classifier artifacts across reporting contexts.

Appendix F. Winsorization sensitivity analysis

This appendix assesses the robustness of the main results to alternative treatments of extreme values in the disclosure-based indices. Specifically, we re-estimate the baseline specifications after winsorizing the dependent variables at increasingly conservative thresholds.

For the Cheap Talk Index, the coefficient on the Biodiversity Financial Risk Score remains positive, stable in magnitude, and statistically significant at the 1% level across all winsorization thresholds (Table A.20).

Table A.20. Robustness of the CRE CTI regression to alternative winsorization thresholds

Winsorization	BRFR score coefficient	Std. error	p-value	R^2
1–99% (baseline)	0.0087***	(0.0025)	0.0006	0.889
2.5–97.5%	0.0085***	(0.0025)	0.0006	0.887
5–95%	0.0078***	(0.0022)	0.0003	0.899

Notes: Results from Model 2 with industry × year and region fixed effects. Equivalent results are obtained for Models 1 and 3. Observations: 896.

For the Positive Talk Index, the coefficient on the Biodiversity Policy Stringency Score likewise remains positive and statistically significant at conventional levels (Table A.21). The stability of the estimates across progressively stricter truncation of the dependent variable indicates that the main findings are not driven by extreme values in the indices.

Table A.21. Robustness of the CRE PTI regression to alternative winsorization thresholds

Winsorization	Policy Stringency Coefficient	Std. error	p-value	R^2
1–99% (Baseline)	0.0085***	(0.0030)	0.0055	0.920
2.5–97.5%	0.0082***	(0.0029)	0.0068	0.923
5–95%	0.0079***	(0.0030)	0.0086	0.925

Notes: Results from Model 2 – Industry × Year and Region FE. Equivalent results are obtained for Models 1 and 3. Observations: 982.

Appendix G. Robustness of the PTI CRE regression to excluding low negative-count observations

The Positive Talk Index is constructed as a log ratio of positive to negative biodiversity-related paragraphs, with a unit constant added to both the numerator and denominator. An inherent feature of this formulation is that firm–year observations with very few negative paragraphs can yield high index values. This is not, in itself, a flaw of the measure: the absence of negative disclosure is a meaningful and conceptually relevant characteristic of corporate reporting behavior, particularly in a context where firms may strategically avoid discussing adverse biodiversity impacts.

At the same time, ratio-based transformations can place greater weight on observations with small denominators. To ensure that the estimated relationship between biodiversity policy stringency and positive talk is not driven by this feature, we conduct a targeted robustness exercise. We re-estimate the baseline CRE Mundlak specification under progressively stricter sample restrictions based on the minimum number of negative biodiversity-related paragraphs required for inclusion. As expected, imposing increasingly strict minimum negative-paragraph thresholds leads to a substantial reduction in the number of firm–year observations, reflecting the empirical rarity of explicit negative biodiversity disclosure in corporate reports over the period studied.

Table A.22 reports the results. Across all specifications, the coefficient on the Biodiversity Policy Stringency Score remains positive and statistically significant, and its magnitude increases as the sample is restricted to firm–years with more substantive negative disclosure. This pattern indicates that the baseline results are not driven by mechanically inflated PTI values arising from zero or near-zero denominators. If anything, the association between policy stringency and positive talk becomes stronger once attention is restricted to reports that meet progressively stricter minimum thresholds for negative biodiversity-related disclosure. These findings provide strong reassurance that the main PTI results reflect systematic differences in disclosure tone rather than artifacts of the index construction.

Table A.22. Robustness of the CRE PTI regression to excluding low negative-count observations

Threshold	Policy Stringency Coefficient	Std. error	p-value	Obs.
No Restriction	0.0085***	(0.0030)	0.0055	982
≥ 1 Negatives	0.0111***	(0.0040)	0.0058	417
≥ 2 Negatives	0.013**	(0.0065)	0.0044	192

Notes: Results from Model 2 – Industry × Year and Region FE. Equivalent results are obtained for Models 1 and 3.

Appendix H. Robustness of event-study results using FY 2020 as baseline

This appendix evaluates the sensitivity of the event-study results to the choice of reference period by re-estimating the correlated random effects Mundlak specifications using FY 2020 as the omitted baseline. Although FY 2020 precedes the Kunming Declaration, sustainability reports for this year were published in 2021 and may already reflect anticipatory disclosure responses. FY 2020 therefore provides a useful but potentially contaminated benchmark for assessing identification.

All specifications replicate the event-study models described in Section 5, with the sole modification being the choice of baseline year.

Appendix H.1. Cheap Talk

Using FY 2020 as the reference period yields strong evidence against parallel pre-trends in cheap talk. All pre-event coefficients (FY 2017–FY 2019) are negative and statistically significant ($\beta_{2017} = -0.094$, $p = 0.004$; $\beta_{2018} = -0.063$, $p = 0.026$; $\beta_{2019} = -0.068$, $p = 0.008$), and a joint Wald test rejects the null of jointly zero pre-baseline coefficients ($\chi^2(3) = 10.00$, $p = 0.019$). This violation indicates that FY 2020 does not constitute a valid pre-event benchmark for identifying post-2020 changes in cheap talk.

Consistent with this identification failure, post-event coefficients are small and statistically insignificant. Estimated effects for FY 2021 ($\beta_{2021} = 0.005$, $p = 0.81$), FY 2022 ($\beta_{2022} = -0.006$, $p = 0.83$), and FY 2023 ($\beta_{2023} = 0.027$, $p = 0.24$) are close to zero and imprecisely estimated, indicating that the increase in cheap talk documented in the main text is mechanically absorbed into the FY 2020 baseline.

EU interaction terms are individually insignificant across all event times, and a joint test of post-2021 EU interaction terms provides no evidence of differential dynamics under this baseline ($\chi^2(3) = 4.78$, $p = 0.19$).

Appendix H.2. Positive Talk

Using FY 2020 as the reference period yields a more nuanced identification pattern for positive talk. In contrast to cheap talk, pre-event coefficients do not provide evidence of systematic divergence in earlier years. Estimates for FY 2017–FY 2019 are small in magnitude and statistically insignificant ($\beta_{2017} = 0.020$, $p = 0.84$; $\beta_{2018} = 0.060$, $p = 0.52$; $\beta_{2019} = -0.022$, $p = 0.77$), and a joint Wald test fails to reject the null of jointly zero pre-baseline coefficients ($\chi^2(3) = 1.41$, $p = 0.70$). Together, these results indicate no detectable violation of parallel pre-trends for positive talk under the FY 2020 baseline.

Post-event coefficients point to a delayed response. Estimated effects for FY 2021 are positive but statistically insignificant ($\beta_{2021} = 0.099$, $p = 0.16$), while effects become positive and statistically significant from FY 2022 onwards ($\beta_{2022} = 0.263$, $p = 0.002$; $\beta_{2023} = 0.315$, $p < 0.001$). This pattern suggests that, unlike cheap talk, positive talk adjustments are not fully absorbed into the FY 2020 baseline and instead emerge with a lag.

EU interaction terms are imprecisely estimated in FY 2021 and FY 2022 ($\delta_{2021} = 0.002$, $p = 0.99$; $\delta_{2022} = -0.198$, $p = 0.15$), but the interaction for FY 2023 is negative and statistically significant ($\delta_{2023} = -0.347$, $p = 0.028$). A joint test of post-event EU interaction terms yields marginally significant evidence of differential dynamics ($\chi^2(3) = 7.40$, $p = 0.060$), suggesting that EU firms exhibit attenuated positive talk responses, with effects loading primarily on FY 2023 under this baseline.

Appendix H.3. Summary

Re-estimating the event-study specifications using FY 2020 as the baseline weakens identification for cheap talk relative to the FY 2019 baseline employed in the main text, as evidenced by strong violations of parallel pre-trends and the absence of post-event effects. For positive talk, parallel pre-trends are satisfied and post-event increases emerge with a delay, consistent with the main specification, although evidence of EU heterogeneity is concentrated in later years under this baseline. Taken together, these results reinforce the choice of FY 2019 as the preferred baseline for jointly interpreting cheap talk and positive talk dynamics around the Kunming declaration,

while showing that the core positive talk findings are robust to using FY 2020 as the reference period.

Appendix I. OLS Robustness Tables

Table A.23 – OLS Cheap Talk Results

Variables	(1)	(2)	(3)
Biodiversity Policy Stringency Score	0.0004 (0.001)	0.0002 (0.0011)	0.0001 (0.001)
Biodiversity Financial Risk Score	0.0062*** (0.002)	0.0071*** (0.0021)	0.0060*** (0.0021)
Environmental Score	0.0013* (0.0008)	0.0016** (0.0008)	0.0014* (0.0008)
log Assets	0.0158 (0.0134)	0.0139 (0.0141)	0.0171 (0.0134)
ROA	0.1477 (0.2335)	0.2066 (0.2506)	0.1497 (0.2423)
Leverage	-0.0899 (0.0833)	-0.0535 (0.0826)	-0.0845 (0.0842)
Years in FG500	0.0004 (0.0013)	0.0007 (0.0014)	0.0003 (0.0013)
Assets Growth	-0.0188 (0.063)	-0.0417 (0.0724)	-0.0319 (0.0685)
Net Income Growth	-0.0023* (0.0014)	-0.0019 (0.0016)	-0.0024* (0.0013)
Num. Biodiversity Paragraphs	-0.0007* (0.0003)	-0.0008** (0.0004)	-0.0007** (0.0004)
Number of Pages	0.0001 (0.0001)	0.0 (0.0001)	0.0001 (0.0001)
GRI	-0.0156 (0.0205)	-0.0194 (0.0215)	-0.0134 (0.0206)
SASB	0.0313 (0.0208)	0.0396* (0.0212)	0.0194 (0.0217)
ISSB	0.0385 (0.026)	0.0332 (0.0269)	0.0407 (0.0277)
ESRS	-0.013 (0.035)	-0.0177 (0.0383)	-0.0099 (0.0382)
TCFD	0.0239 (0.0187)	0.0279 (0.0193)	0.0336* (0.0192)
TNFD	-0.0101 (0.0371)	0.0007 (0.0366)	0.0007 (0.0383)
Assurance Profile Big 4 Comprehensive	0.0232 (0.0231)	0.0197 (0.0243)	0.0201 (0.0234)
Assurance Profile Big 4 Limited	0.0512 (0.0365)	0.043 (0.0389)	0.0495 (0.0373)
Assurance Profile Other Comprehensive	0.0679 (0.0701)	0.0671 (0.0645)	0.0535 (0.0711)
Assurance Profile Other Limited	-0.0761 (0.0474)	-0.0972* (0.0518)	-0.0778 (0.0475)
Year_2018	0.004 (0.0257)		
Year_2019	0.0096 (0.027)		
Year_2020	0.0487 (0.0305)		
Year_2021	0.0367 (0.029)		
Year_2022	0.0274 (0.0327)		
Year_2023	0.0697** (0.032)		
Intercept	0.1712 (0.1575)	0.2497 (0.1709)	0.2161 (0.1588)
FE industry	YES	NO	YES
FE industry × year	NO	YES	NO
FE region	YES	YES	NO
FE region × year	NO	NO	YES
R-squared	0.14	0.21	0.16

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Obs: 896

Table A.24 – OLS PTI Results

Variables	(1)	(2)	(3)
Biodiversity Policy Stringency Score	0.0063** (0.0031)	0.0061* (0.0032)	0.0064** (0.0032)
Biodiversity Financial Risk Score	-0.0034 (0.0084)	-0.0051 (0.009)	-0.0033 (0.0086)
Environmental Score	0.0004 (0.0021)	0.0006 (0.0023)	0.0006 (0.0022)
log Assets	-0.0508 (0.0366)	-0.0501 (0.0377)	-0.0516 (0.0369)
ROA	-0.2154 (0.5975)	-0.0106 (0.6549)	-0.1261 (0.6012)
Leverage	-0.0116 (0.2343)	-0.027 (0.2451)	-0.0165 (0.2352)
Years in FG500	-0.0034 (0.0041)	-0.0036 (0.0043)	-0.0035 (0.0042)
Assets Growth	-0.0783 (0.1728)	-0.0934 (0.2019)	-0.1422 (0.1905)
Net Income Growth	0.0106 (0.0082)	0.0113 (0.0082)	0.0111 (0.0085)
Num. Biodiversity Paragraphs	0.0120*** (0.0014)	0.0118*** (0.0015)	0.0121*** (0.0015)
Number of Pages	0.0005 (0.0004)	0.0005 (0.0004)	0.0005 (0.0004)
GRI	-0.1399** (0.0595)	-0.1345** (0.0615)	-0.1371** (0.0614)
SASB	-0.0899 (0.07)	-0.108 (0.0733)	-0.0911 (0.0724)
ISSB	-0.0968 (0.0989)	-0.0652 (0.106)	-0.0932 (0.1002)
ESRS	-0.1421 (0.1267)	-0.1241 (0.1287)	-0.0866 (0.1383)
TCFD	0.0566 (0.0604)	0.0542 (0.0653)	0.0579 (0.0607)
TNFD	-0.1551 (0.1151)	-0.1328 (0.1226)	-0.1605 (0.1221)
Assurance Big 4 Comprehensive	-0.0364 (0.0786)	-0.0401 (0.0814)	-0.033 (0.0794)
Assurance Big 4 Limited	-0.117 (0.1053)	-0.1342 (0.1114)	-0.1118 (0.1069)
Assurance Other Comprehensive	-0.2842*** (0.1001)	-0.3124*** (0.1067)	-0.2770*** (0.1011)
Assurance Other Limited	0.0782 (0.1147)	0.0888 (0.1274)	0.093 (0.1149)
Year 2018	0.0752 (0.0622)		
Year 2019	-0.0445 (0.0759)		
Year 2020	-0.0011 (0.0875)		
Year 2021	0.0365 (0.0848)		
Year 2022	0.1326 (0.0818)		
Year 2023	0.1406 (0.0855)		
FE industry	YES	NO	YES
FE industry × year	NO	YES	NO
FE region	YES	YES	NO
FE region × year	NO	NO	YES
R-squared	0.36	0.41	0.37

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Obs: 982.

Appendix J. Construct Validation of the Cheap Talk Index

To further strengthen the construct validity of the CTI, we conducted a targeted report-level validation exercise. We drew a stratified random sample of firm–year observations from the top decile of the CTI distribution, ensuring balanced coverage across regions, industries, and years. For each sampled firm–year, all biodiversity-related commitment paragraphs classified by the model were subjected to a blinded manual review. Two researchers

independently labeled each paragraph as specific or non-specific following the annotation guidelines described in Appendix B. Inter-rater agreement was high (Cohen's $\kappa \approx 0.78$), indicating substantial consistency in the human assessment of specificity.

The validation sample comprises 193 paragraphs classified by the model as non-specific commitments. Manual review confirms that 89% of these paragraphs are indeed non-specific, providing strong reassurance that high CTI values reflect a genuine reliance on vague, non-verifiable commitment language. Practically all misclassified paragraphs correspond to boundary cases, in which a concrete initiative or planned action is mentioned, but with limited detail regarding scope, implementation, targets, or monitoring. These cases lie close to the conceptual boundary between specific and non-specific commitments rather than reflecting systematic classification error. This error structure is fully consistent with the patterns documented in the human audit of the specificity classifier reported in Appendix D.

Overall, this exercise supports the interpretation of the Cheap Talk Index as a meaningful measure of firms' reliance on vague biodiversity-related commitments. To illustrate the qualitative patterns underlying high CTI values, we report anonymised and paraphrased examples of non-specific commitment language below:

- *Tackling nature-related impacts and biodiversity loss is a key priority. We have taken a number of early steps in this still emerging and complex area to deepen our understanding and to better manage our effects on nature and biodiversity.*
- *Over the years, we have backed projects and initiatives led by a range of partners to promote forest conservation and reforestation.*
- *We put in place robust assessment processes to identify and monitor risks linked to the sourcing of wood and wood products, while exploring ways to reduce or prevent impacts on biodiversity.*
- *We are continuing dialogue around our key raw materials, building on established foundations to develop biodiversity-specific commitments and targets informed by ongoing, evidence-based strategies.*
- *We will seek to establish collaborative relationships with governments, international institutions, NPOs, suppliers, customers, and a wide range of other groups that work to protect biodiversity.*