SentiPromiseESG: Sentiment Analysis of Sustainability Promises Across Industries

Wei-Chen Huang[1]
wesleyseawatch@gmail.com

Hsin-Ting Lu[2]
hsintinglubob@gmail.com

Wen-Ze Chen[2]
50712andy@gmail.com

Yu-Han Huang[3]
5monica.cupcake@gmail.com

Min-Yuh Day[2]*
myday@gm.ntpu.edu.tw

Department of Statistics[1], Graduate Institute of Information Management[2]
, Department of Accounting[3]
National Taipei University, New Taipei City, Taiwan

Abstract

Sustainability reporting has become a primary channel for firms to communicate their environmental, social, and governance (ESG) performance, yet the voluntary and narrative nature of these disclosures raises persistent concerns about selective reporting and greenwashing. Despite the growing use of large language models (LLMs) and ESG-focused language models to process sustainability texts, there is still limited empirical evidence on how the tone of ESG disclosures relates to the presence, content, and verifiability of sustainability promises, particularly in non-English, industry-specific settings. This study aims to examine how sentiment is associated with sustainability promise behavior by developing a sentiment-aware, LLM-based framework for identifying and evaluating sustainability promises across industries. Using ESG reports from 29 large Taiwanese listed firms in the semiconductor, financial, and computer-peripherals sectors, we construct the SentiPromiseESG Dataset with 15,345 sentence-level spans annotated for promise status, evidence status, evidence quality, verification timeline, and ESG type via a two-stage Gemini 2.5 Pro pipeline, and derive weighted sentiment scores from Gemini 2.5 Pro and FinBERT-ESG that are incorporated into span-level logistic regression models. The empirical results show that more positive sentiment is consistently associated with a higher likelihood of making sustainability promises and with a stronger tendency for those commitments to fall within the social dimension rather than environmental or governance domains, whereas sentiment exhibits only weak and unstable links to whether supporting evidence is provided and how clearly that evidence is presented. These findings contribute a scalable LLM-based annotation and sentiment-analysis pipeline, introduce the SentiPromiseESG Dataset as a new resource for ESG and greenwashing research, and offer practical implications for investors, regulators, and practitioners seeking to use textual analytics and LLM-based tools to assess the credibility and narrative structure of corporate sustainability communication.

Keywords:
Sustainability Promises, Sentiment Analysis , ESG, Greenwashing, Large Language Model

I. INTRODUCTION

Sustainability reporting has become a central channel for communicating corporate priorities and responding to increasing expectations from regulators, investors, and stakeholders. Environmental, social, and governance (ESG) disclosures now serve as key sources for evaluating how firms manage climate risks, labor conditions, corporate governance structures, and long-term sustainability strategies (KPMG, 2022). These reports provide insights not only into past performance but also into forward-looking commitments. Yet because much of ESG reporting remains voluntary, firms retain substantial discretion over what to disclose and how to frame it, enabling selective emphasis and aspirational narratives that may overstate sustainability efforts (Lyon & Montgomery, 2015).

Greenwashing has therefore become a persistent concern in sustainability research. Prior work shows that firms often use symbolic commitments, optimistic tone, or promotional language to project responsibility without implementing corresponding operational changes (Delmas & Burbano, 2011). However, such symbolic or promotional sustainability claims frequently lack measurable indicators, traceable evidence, or verification timelines, which are necessary for credible external evaluation. As a result, unsubstantiated sustainability claims undermine their credibility and complicate external evaluation (Lublóy et al., 2025).

Advances in natural language processing (NLP) and large language models (LLMs) provide new opportunities to address these challenges. General-purpose models such as GPT and Gemini, along with domain-specific models such as FinBERT (Araci, 2019) and ClimateBERT, have been applied to ESG topic classification, sentiment detection, and automated greenwashing analysis. However, ESG texts are heterogeneous, lengthy, and often multi-lingual, making fine-grained information extraction and promise verification non-trivial. Model performance can vary substantially depending on task formulation and prompt design, and LLMs may be misled by repeated keywords or loosely related text spans (Barbeito-Caamaño & Chalmeta, 2020).

At the same time, sentiment analysis has become an important lens for studying corporate communication in finance and ESG. Prior studies construct tone indicators for environmental, social, and governance dimensions and link them to outcomes such as risk, valuation, or disclosure quality. In ESG contexts, sentiment-based indicators have been used to measure reporting quality and environmental messaging strategies (Barbeito-Caamaño & Chalmeta, 2020). Yet few studies examine how sentiment interacts directly with sustainability promises themselves—whether positive tone increases the likelihood of making commitments, influences evidence provision, or shifts the distribution of ESG promise types. Empirical work integrating sentiment modelling and automated promise verification remains especially limited in non-English and industry-specific settings.

To address these gaps, we propose a sentiment-aware analytical framework for identifying and evaluating sustainability promises in ESG reports. Focusing on large Taiwanese listed firms across the semiconductor, financial, and computer-peripherals sectors, we construct the SentiPromiseESG Dataset using a two-stage LLM-based annotation pipeline. We employ Gemini 2.5 Pro to extract candidate sustainability-related spans and assign labels—including promise status, evidence status, evidence quality, verification timeline, and ESG type—under a 5-shot prompting setup derived from ML-Promise (Seki et al., 2024). Sentiment scores are sourced from Gemini 2.5 Pro and FinBERT-ESG and mapped to continuous polarity indices.

Building on this dataset, we examine four research questions using logistic regression models at the text-span level:

1. Is sentiment associated with the presence of sustainability promises?
2. Is sentiment associated with the provision of verifiable evidence?
3. Is sentiment associated with the distribution of ESG promise types (social vs. environmental, governance vs. environmental)?

4. Is the sentiment of evidence text associated with the clarity of the evidence?

The contributions of this paper are threefold. First, we propose a scalable, LLM-based pipeline for automatic sustainability promise annotation that integrates dual-model sentiment scoring, addressing the need for more robust tools for ESG text analysis. Second, we provide a large-scale empirical design to study the relationship between sentiment tone, promise presence, evidence provision, and ESG-type distributions across industries. Third, we introduce the SentiPromiseESG Dataset as a new resource for research on greenwashing, report quality, and automated evaluation of sustainability communication.

The remainder of this paper is organized as follows. Section II reviews related work on ESG disclosure, greenwashing, LLM-based ESG text analysis, and commitment and sentiment modelling. Section III describes the construction of the SentiPromiseESG Dataset, the sentiment scoring procedure, and the regression design. Section IV presents the empirical results and discusses the four research questions. Section V concludes with a summary of the findings, implications, limitations, and directions for future research.

II. LITERATURE REVIEW

2.1 The Rise of ESG Disclosure and Greenwashing Practices

As sustainability has become a central focus in corporate governance, ESG reports have gradually become a primary source of information for stakeholders to assess corporate sustainability performance (Kim et al., 2023). Sustainability reports allow external audiences to understand a company's strategies and outcomes across environmental, social, and governance dimensions. However, because most sustainability information is disclosed voluntarily, firms possess considerable discretion when preparing such reports. This discretion may lead to selective presentation of favorable information and the downplaying of negative aspects, which creates information asymmetry and reduces the objectivity and completeness of ESG disclosures (Xu et al., 2025).

Against this backdrop, greenwashing has drawn increasing attention. Greenwashing refers to the use of misleading narratives, vague language, or symbolic actions to enhance a company's environmental image and create the false impression that its sustainability performance is stronger than it actually is (He & Wang, 2025). Prior research indicates that companies may emphasize visions or slogan-like commitments to construct an impression of sustainability. Yet such commitments frequently lack measurable indicators, verification timelines, or traceable evidence. This weakens the credibility of sustainability reports and makes it difficult for external evaluators to determine whether corporate commitments are genuine or feasible.

Therefore, an important topic in ESG research is how to identify, within extensive sustainability texts, whether companies provide specific and verifiable sustainability commitments and whether their narratives contain substantive content.

2.2 Large Language Model Driven ESG Text Analysis and its Challenges

In recent years, large language models, including OpenAI's GPT series, Google's Gemini, and domain specific models such as FinBERT and ClimateBERT, have been widely applied in ESG text analysis. These applications include sentiment classification, identification of corporate commitments, and detection of greenwashing practices (Birti et al., 2025). Despite these advances, the use of large language models in ESG related corpora faces several challenges. General purpose models have limited ability to capture fine grained ESG semantics and often require high quality domain specific data for adaptation, yet such training resources remain scarce. Cross linguistic analysis is also difficult because ESG reports across languages contain diverse and extensive content, which makes automated commitment verification highly challenging (Turk et al., 2025). Furthermore, ESG reports are typically lengthy and stylistically inconsistent, increasing the difficulty of achieving reliable and coherent model interpretation.

Existing studies show that when models lack explicit guidance, the accuracy of extracting information from unstructured sustainability reports can fall below 30 percent. However, when provided with specific KPI related prompts and standardized input formats, the accuracy can exceed 70 percent (Martín-Domingo et al., 2025). Even with state of the art models, substantial obstacles remain when processing long documents that span multiple languages or inconsistent contexts. Examples include interference from irrelevant information, ambiguous language usage, and variation in stakeholder perspectives (Ong et al., 2025). These limitations highlight the continued need to strengthen multilingual support, improve contextual understanding, and incorporate domain knowledge in ESG text analysis, which remain essential directions for future research.

2.3 Applications of Natural Language Processing in Commitment Identification and Sentiment Analysis in ESG Texts

Natural language processing has been applied to the identification of commitment statements and sentiment analysis in sustainability reports. Existing approaches include keyword based rule retrieval, fine tuning BERT models to classify commitment sentences, and using GPT with few shot prompting for information extraction (Schimanski et al., 2024). In sentiment analysis, one study constructed tone indicators for environmental, social, and governance dimensions and reported that positive tone in certain dimensions is significantly associated with lower risk profiles. However, limitations persist due to semantic ambiguity and misclassification. Traditional NLP techniques have difficulty understanding contextual meaning. Because of repeated or ambiguous keywords, algorithms often struggle to determine which matched segment is most relevant when the same term appears in different contexts within a report (Sun et al., 2024). This may result in extracted outputs that include information not present in the report or fail to capture information that is actually mentioned.

Furthermore, models generally lack the ability to incorporate external evidence. When companies enhance their ESG performance through favorable wording, a phenomenon associated with greenwashing, purely text based analysis can be easily misled. These limitations highlight the need to improve models for assessing the credibility of commitments and predicting tonal characteristics, and they provide the motivation for subsequent research in this area.

2.4 Summary

Overall, this chapter provides an in-depth examination of the key challenges associated with ESG information disclosure in the context of corporate sustainability governance. As sustainability reporting becomes more common, the voluntary nature of disclosure allows firms to selectively present favorable information and employ vague language or commitments that lack specific indicators, which contributes to greenwashing and undermines the credibility of the reports. To address these issues, large language models have been widely applied to ESG text analysis, yet they continue to face challenges such as limited ability to recognize fine grained semantics, scarcity of domain specific data, and difficulty in understanding long and complex documents. Although existing NLP techniques can identify commitment statements and analyze sentiment, their ability to interpret contextual meaning and incorporate external evidence remains limited, making it difficult to reliably detect greenwashing practices. Therefore, the motivation of this study is to address this critical gap by using more advanced models to efficiently and accurately assess the credibility of corporate commitments within extensive sustainability texts.

III. METHODOLOGY

3.1 Research Framework and Workflow

This study adopts an end-to-end workflow that links promise extraction, sentiment data acquisition, and regression analysis, as summarized in Figure 1.

In the promise data extraction and annotation stage, ESG reports from Yuanta 50 constituent firms are processed by Gemini 2.5 Pro. The model scans each report, extracts candidate sustainability-related text spans, and assigns PromiseEval-style labels (promise status, evidence status, evidence quality, verification timeline, and ESG type) under a 5-shot prompting setup using examples adapted from the ML-Promise Japanese dataset. The resulting JSON files constitute the SentiPromiseESG annotated corpus.

In the sentiment data acquisition stage, each annotated span is translated into English where necessary using Gemini 2.5 Pro and then evaluated by two models: Gemini 2.5 Pro and FinBERT-ESG. Both models output probabilities for positive, neutral, and negative classes, which are transformed into continuous polarity indices and combined into weighted sentiment scores for statements and evidence text.

In the regression analysis stage, these weighted sentiment scores are merged with the annotation labels and used as key explanatory variables in a set of logistic regression models. The models are estimated at the text-span level to examine how sentiment is associated with promise presence, evidence provision, ESG-type classifications, and evidence clarity across firms and industries.

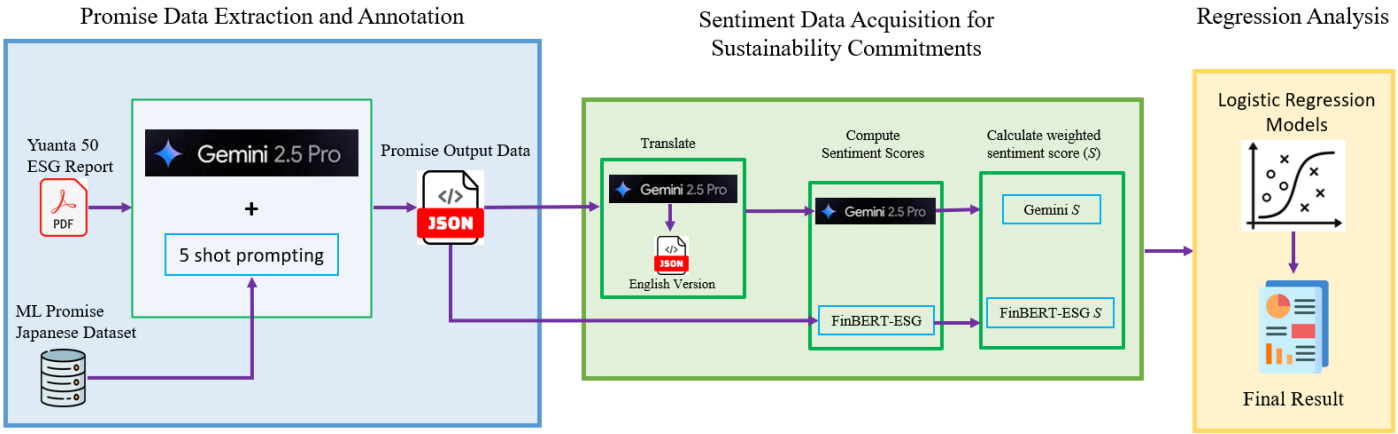


Figure1: Overall Workflow of the SentiPromiseESG Analysis Framework
Source：This study

3.2 Construction of the Sustainability Promise Annotation Dataset

Building on the research framework in Section 3.1, the empirical analysis is based on ESG disclosures from large Taiwanese listed firms. In the first step, we selected candidate companies from the Yuanta Taiwan 50 index, which comprises the 50 largest firms by market capitalization. Using the industry classifications provided by Yahoo Finance (Taiwan), we identified financial services, computer peripherals, and semiconductors as the three most represented sectors in the index. We then restricted our sample to firms in these three industries and collected their latest sustainability (ESG) reports, yielding 29 companies that broadly represent Taiwan's dominant sectors in terms of market capitalization and ESG communication.

Building on these reports, we constructed a sentence-level sustainability-promise dataset using a two-stage LLM-based pipeline. In the first stage, Gemini 2.5 Pro was prompted to scan each PDF report (after image-based rendering where necessary) and extract candidate text spans likely to contain sustainability-related statements. In the second stage, each extracted span was passed to Gemini 2.5 Pro under a 5-shot prompting setting to assign promise-related labels. We

adapt in-context examples from the Japanese portion of the ML-Promise dataset (Seki et al., 2024), released as part of the SemEval-2025 Task 6 "PromiseEval" shared task on corporate promise verification. Leveraging Gemini's multilingual capabilities, we reuse these Japanese exemplars as in-context demonstrations when prompting on Chinese ESG texts.

We choose Gemini 2.5 Pro because it supports joint text–vision reasoning on PDF page images and currently ranks among the top models in both the "Text" (pure text) and "Vision" (image+text) tracks on the LMArena leaderboard hosted on Hugging Face (LMArena, 2025).

The annotation schema follows PromiseEval with minor adaptations to the ESG context. For each text span we record:

1. **Promise Status** – whether a concrete or organization-level sustainability commitment is present (Yes/No).
2. **Evidence Status** – whether verifiable supporting evidence is provided (Yes/No).
3. **Evidence Quality** – clarity of the supporting evidence (Clear, Not Clear, Misleading, N/A).
4. **Verification Timeline** – expected time frame for fulfilling the promise (Already, Within 2 years, Between 2–5 years, More than 5 years, N/A).
5. **ESG Type** – topical category of the statement (Environment, Social, Governance).

In particular, we do not remove N/A values. When Evidence Status is "No," the corresponding Evidence Quality is coded as N/A; dropping all N/A values for Evidence Quality would therefore eliminate all instances with Evidence Status = "No." The same rationale applies to Verification Timeline: when Promise Status is "No," Verification Timeline is recorded as N/A, and these observations are retained in the dataset.

For each span, the resulting JSON structure includes the original text (data), the extracted promise clause (promise_string), the evidence clause (evidence_string), and all associated categorical labels. Aggregating across the 29 firms, this procedure yields 15,345 annotated text spans, as reported in Table 1, which constitute the SentiPromiseESG dataset used in the subsequent analyses.

Task	Label	Count
Promise Status	Yes	7251 (47.25%)
	No	8094 (52.75%)
Evidence Status	Yes	5591 (36.44%)
	No	1988 (12.96%)
	N/A	7766 (50.61%)
Evidence Quality	Clear	4455 (29.03%)
	Not Clear	1106 (7.21%)
	Misleading	30 (0.20%)
	N/A	9754 (63.56%)
Verification Timeline	Already	3549 (23.13%)
	Within 2 years	2320 (15.12%)
	Between 2 to 5 years	463 (3.02%)
	More than 5 years	918 (5.98%)
	N/A	8095 (52.75%)
ESG Type	E	4027 (26.24%)
	S	6711 (43.73%)
	G	4607 (30.02%)

Table 1: Label distribution of the SentiPromiseESG Dataset

3.3 Sentiment Scoring of Sustainability Promises

To quantify the tone of sustainability communication, we derive sentiment scores for three textual fields: the original ESG statement (data), the extracted promise string, and the evidence

string. Because our texts are primarily in Chinese, we first use Gemini 2.5 Pro to translate each segment into English while preserving domain-specific terminology.

For sentiment modelling, we combine a general-purpose multilingual LLM, Gemini 2.5 Pro, with a domain-specialized model, FinBERT-ESG. FinBERT is a BERT-based architecture pre-trained and fine-tuned on financial texts for sentiment classification and has been shown to achieve state-of-the-art performance on several finance sentiment benchmarks (Araci, 2019). We use the publicly available FinBERT-ESG checkpoint released on Hugging Face (Huang et al., 2023), which is further fine-tuned on ESG-related disclosures and is therefore better aligned with our sustainability-report setting. FinBERT-ESG has also been employed in recent empirical ESG studies using sustainability reports and stock-market data (Atak, 2024). In line with the sentiment analysis and opinion-mining literature, which typically represents affect along discrete polarities (positive, neutral, negative) and then aggregates them into a single index, we obtain, for each segment, three probabilities of positive, neutral and negative from FinBERT-ESG, applied to the English translations, and from Gemini 2.5 Pro, directly queried for probabilistic sentiment judgments. Both models output probabilities that lie between 0 and 1 and sum to 1 across the three sentiment classes.

Descriptive statistics of the resulting sentiment distributions are reported in Table 2 and Table 3. Gemini assigns relatively higher average probabilities to the neutral class across all three text fields, whereas FinBERT-ESG produces more polarized outputs with higher mean scores for the positive class.

Following common practice in financial and social-media sentiment research, where discrete class probabilities are mapped to a continuous polarity score (Liu, 2022), we adopt a simple but interpretable extreme-value coding scheme: positive = +1, neutral = 0, negative = −1. For each segment we compute a **weighted sentiment score :**

$$S = 1 \cdot p_{pos} + 0 \cdot p_{neu} - 1 \cdot p_{neg}$$

which yields a continuous index in $[-1,1]$, where higher values indicate more positive affect and lower values indicate more negative affect. This transformation is applied to both FinBERT-ESG and Gemini outputs; in the subsequent regression analyses we use the combined (averaged) score as our main predictor for the statement-level tone, and the evidence-specific score as a predictor for evidence clarity.

This design has two advantages. First, it allows us to exploit the complementary strengths of a multilingual LLM and a finance-specific model while keeping the downstream representation one-dimensional and interpretable. Table 4 provides an illustrative example from the SentiPromiseESG Dataset, showing how each annotated text span includes the translated statement, extracted promise and evidence clauses, and the corresponding sentiment probabilities used to construct the weighted sentiment scores.

Category	Variable	Count	Mean	Std
Data Sentiment	Negative	15345	0.0744	0.1299
	Neutral	15345	0.6625	0.2320
	Positive	15345	0.2632	0.2331
Promise Sentiment	Negative	7251	0.0425	0.0846
	Neutral	7251	0.6456	0.2398
	Positive	7251	0.3118	0.2495
Evidence Sentiment	Negative	5591	0.0564	0.1224
	Neutral	5591	0.7025	0.2289
	Positive	5591	0.2411	0.2283

Table 2: Descriptive statistics of sentiment scores for the SentiPromiseESG dataset (Gemini 2.5 Pro)

Category	Variable	Count	Mean	Std
Data Sentiment	Negative	15345	0.1966	0.3273
	Neutral	15345	0.2789	0.4056
	Positive	15345	0.4044	0.4145
Promise Sentiment	Negative	7251	0.1075	0.2412
	Neutral	7251	0.3269	0.4342
	Positive	7251	0.4370	0.4290
Evidence Sentiment	Negative	5591	0.1609	0.2966
	Neutral	5591	0.3043	0.4244
	Positive	5591	0.4218	0.4230

Table 3: Descriptive statistics of sentiment scores for the SentiPromiseESG dataset (FinBERT-ESG)

Field	Value
Data Translated	The UMC Group has established a stakeholder engagement mechanism to identify key stakeholders and regularly disclose information on topics of concern through various communication channels, effectively communicating with stakeholders. Goals: Understand the reasonable expectations and needs of stakeholders , and appropriately respond to their key ESG issues of concern. Consider all issues of concern and analyze potential environmental, social, economic, and operational impacts on the company. Continuously review and improve through a systematic mechanism to enhance sustainability performance.
URL	https://www.umc.com/upload/media/07_Sustainability/72_Reports_and_Results/ 1_Corporate_Sustainability_Reports/CSR_Reports/CS_Reort_chinese_pdf/ 2024_CSR_report_chi/UMC-2024-CH-elink.pdf
Page Number	19
ESG Type	G
Promise Status	Yes
Promise String Translated	Understand the reasonable expectations and needs of stakeholders and appropriately respond to significant ESG issues they are concerned about. Consider all relevant issues and analyze their potential environmental , social, economic, and operational impacts. Continuously review and improve through systematic mechanisms to enhance sustainability performance
Verification Timeline	already
Evidence Status	Yes
Evidence String Translated	The United Microelectronics Corporation (UMC) Group has established a stakeholder engagement mechanism to identify key stakeholders and continuously disclose information on issues of concern through various communication channels, effectively engaging with stakeholders.
Evidence Quality	Not Clear
Data Sentiment	"negative": 0.0053, "neutral": 0.6799, "positive": 0.2967
Promise Sentiment	"negative": 0.0149, "neutral": 0.894, "positive": 0.0822
Evidence Sentiment	"negative": 0.0063, "neutral": 0.0066, "positive": 0.5481

Table 4: Example entry from the SentiPromiseESG Dataset

3.4 Regression Modelling

To examine whether the sentiment of sustainability disclosures is systematically associated with promise behavior and evidence provision, we estimate a series of logistic regression models at the text-span level. Logistic regression is well-suited for modelling binary outcomes and is widely used in the social sciences and applied statistics. For each research question, we specify the following dependent variables:

1. **Promise status** – a binary indicator of whether a span contains a concrete sustainability promise (Promise Status = Yes vs. No).
2. **Evidence status** – a binary indicator of whether verifiable evidence is provided (Evidence Status = Yes vs. No, excluding N/A).

3. **ESG type** – two binary contrasts comparing Social vs. Environmental (ESG_S = 1 if S, 0 if E) and Governance vs. Environmental (ESG_G = 1 if G, 0 if E), respectively.
4. **Evidence quality** – a binary indicator of whether the evidence is assessed as clear (Evidence Quality = Clear vs. any other category).

The key independent variables are the **weighted statement-level sentiment score** from data (S_data) and the **weighted evidence-level sentiment score from evidence(**S_ evidence**)**, derived as described in Section 3.2. For the promise-, evidence-, and ESG-type models, we use S_data as the focal predictor; for the evidence-quality model, we use S_ evidence as the main predictor.

All models are estimated using maximum likelihood. We report logit coefficients β, their associated odds ratios (OR = e^{β}), 95% confidence intervals, and p-values. To guard against heteroskedasticity and model misspecification, we compute heteroskedasticity-consistent (HC1) robust standard errors (White, 1980). Model calibration and discriminative ability are assessed using likelihood-ratio statistics, pseudo-R^2 measures, and the area under the ROC curve (AUC), in line with standard recommendations for applied logistic regression.

To address industry heterogeneity, we estimate each set of models twice: once on the pooled dataset across all firms, and once separately for the three major sectors (semiconductors, financials, and computer peripherals). This design allows us to test not only whether sentiment scores are associated with sustainability promises and evidence at the aggregate level, but also whether these relationships differ systematically across industries with distinct business models and regulatory environments.

Figure 1 summarizes this methodological design by linking the stages of data collection, LLM-based annotation, sentiment scoring, and regression modelling in a single end-to-end pipeline.

IV. RESULTS AND ANALYSIS

This section draws on the empirical results from the SentiPromiseESG Dataset to examine how sentiment scores are related to firms' sustainability commitment behavior (ESG promises). The aggregate regression results are reported in Tables 4 and 6, while the industry-specific estimations are presented in Tables 5 and 7. All regression models employ HC1 heteroscedasticity-consistent standard errors to correct for heteroskedasticity and ensure the robustness of the inference. Model performance evaluations indicate that the AUC of all specifications exceeds 0.6, suggesting that the sentiment scores possess non-trivial predictive power. The discussion is organized around the four research questions: (1) whether sentiment affects the presence of a promise, (2) whether it affects the provision of supporting evidence, (3) whether it shifts the distribution of ESG promise types, and (4) whether the sentiment of the evidence affects its clarity. For each question, we compare the results from the full sample and the industry-level estimations.

4.1 Sentiment Scores and the Existence of Sustainability Promises

For the full sample, Table 5 shows that when using the Gemini-generated weighted document sentiment score (DSW), the coefficient in the Promise model is $\beta = 0.9767$ with an odds ratio (OR) of 2.656, statistically significant at the 0.1% level ($p < .001$). This implies that a one-unit increase in the sentiment score is associated with roughly a 2.7-fold increase in the odds that a text passage contains at least one sustainability promise.

By comparison, Table 7 reports a smaller coefficient for FinBERT-ESG is $\beta = 0.3626$, but the corresponding OR remains above one at 1.437, again significant at the 0.1% level. Thus, under both sentiment models, we obtain a consistent positive association between more positive sentiment and the likelihood that a promise is present.

At the industry level, Table 6 indicates that the Gemini-based Promise models yield ORs between 2.46 and 3.00 across the three industries, with coefficients statistically significant at the 1% level or better. This suggests that in the semiconductor, financial, and computer-peripherals industries alike, more positive sentiment is systematically associated with a higher probability that promise s are made. As shown in Table 8, the FinBERT-ESG Promise models likewise point in the same direction: the ORs range from approximately 1.25 to 1.56 across the three industries, and all coefficients are significant at the 5% level or better, further reinforcing the positive link between sentiment and promise presence.

Taken together, the full-sample and industry-specific results suggest that positive sentiment persistently and significantly increases the likelihood that sustainability promises are put forward. This pattern is robust across models (Gemini and FinBERT-ESG) and across industries, indicating a clear and consistent relationship between promise behavior and the sentiment of the underlying text.

4.2 Sentiment Scores and the Provision of Supporting Evidence

For the full sample, the two sentiment models yield different conclusions regarding the provision of evidence. In Table 5, the Gemini Evidence Status model reports a DSW coefficient of $\beta = -0.4667$ with an OR of 0.627, significant at the 0.1% level ($p < .001$). Under this specification, more positive sentiment is associated with lower odds that the text is labeled as "with evidence." In contrast, the FinBERT-ESG Evidence Status model in Table 7 produces a much smaller coefficient ($\beta = -0.0341$), with an OR close to 1 (0.967) and a p-value of .464, which is not significant at the 5% level. Under this alternative sentiment measurement, there is no stable linear relationship between sentiment and the provision of evidence.

At the industry level, the Gemini results in Table 6 show that the ORs in the Evidence Status models for the semiconductor and financial industries lie between 0.60 and 0.63, with coefficients statistically significant at the 1% level. For the computer-peripherals industry, however, the OR is around 0.83 and does not pass the 5% significance threshold, suggesting that the negative association does not generalize to all industries. Turning to Table 8, the FinBERT-ESG Evidence Status models exhibit ORs between approximately 0.97 and 1.17 across the three industries, with none of the coefficients statistically significant at the 5% level. Under the FinBERT-ESG sentiment measure, sentiment scores do not exert a systematic impact on whether evidence is provided.

Overall, although the Gemini model and some industries exhibit a pattern whereby more positive language is associated with a lower likelihood of providing evidence, this result cannot be consistently replicated across both sentiment models and all industries. Consequently, we adopt a cautious interpretation regarding the hypothesis that sentiment systematically reduces the probability of evidence provision. The overall impact appears limited and subject to heterogeneity across models and industries.

4.3 Sentiment Scores and the Distribution of ESG Promise Types

For the full sample, Table 5 shows that in the Gemini S vs. E model, the coefficient of DSW is $\beta = 1.4258$ with an OR of 4.161 ($p < .001$), while in the G vs. E model the coefficient is $\beta = -1.4917$ with an OR of 0.225 ($p < .001$). Table 7 reports qualitatively similar patterns for FinBERT-ESG: in the S vs. E model, the coefficient is $\beta = 1.5173$ with an OR of 4.560 ($p < .001$), whereas in the G vs. E model it is $\beta = -0.5849$ with an OR of 0.557 ($p < .001$). In other words, as sentiment becomes more positive, a text passage is more likely to be labeled as belonging to the social (S) rather than the environmental (E) dimension, and less likely to be labeled as governance (G) rather than environmental (E).

At the industry level, Table 6 indicates that for the Gemini S vs. E models, the ORs across the three industries range from 2.62 to 5.74, with all coefficients significant at the 1% level. In

the G vs. E models, the ORs fall between 0.18 and 0.29, again all significant at the 1% level. Table 8 reveals comparable patterns for the FinBERT-ESG models: in the S vs. E specifications, the ORs range from about 3.39 to 4.76, with all industry coefficients significant at the 1% level, implying that more positive sentiment makes texts more likely to be classified as social rather than environmental. In the G vs. E models, the ORs lie between roughly 0.47 and 0.67, also uniformly significant at the 1% level, indicating that in more positive passages, governance-related commitments are less likely to appear relative to environmental ones.

Taken together, the full-sample and industry-specific analyses reveal that positive sentiment is strongly and consistently positively associated with "social-dimension commitments" and negatively associated with "governance-dimension commitments." This regularity holds across sentiment models and industries, suggesting that ESG categories exhibit distinct narrative styles that correlate with systematic differences in the tone of the language.

4.4 Evidence Sentiment and the Clarity of Evidence

The fourth research question examines whether the sentiment of the evidence text affects the clarity with which the evidence is articulated. For the full sample, Table 5 shows that in the Gemini Evidence Quality model, the coefficient on the weighted evidence sentiment score (ESW) is $\beta = 0.0199$ with an OR of 1.020 and $p = .866$, which is not statistically significant. Thus, the Gemini-based sentiment measure does not support the hypothesis that more positive sentiment leads to clearer evidence.

By contrast, the FinBERT-ESG results in Table 7 report an ESW coefficient of $\beta = 0.1399$ with an OR of 1.150 and $p = .014$. Although the magnitude of the effect is modest, it is statistically significant at the 5% level. Under the FinBERT-ESG model, more positive evidence text is thus mildly associated with a higher probability of being labeled as clearer evidence.

From an industry perspective, Table 6 shows that in the Gemini industry-level Evidence Quality models, all three industries have ORs close to 1, and none of the coefficients are statistically significant at the 5% level. In other words, under the Gemini measure, there is no detectable statistical relationship between evidence sentiment and evidence clarity in any of the three industries. Table 8 presents the FinBERT-ESG industry-level Evidence Quality models, where the ORs range from about 1.08 to 1.20. Among these, only the coefficient for the financial industry reaches significance at the 5% level; the coefficients for the other industries are not statistically significant. This pattern suggests that even under FinBERT-ESG, the notion that "more positive evidence is more likely to be labeled as clear" appears to be concentrated in specific industries rather than a universal phenomenon.

Overall, the influence of evidence sentiment on evidence clarity is relatively weak and unstable. In most cases, whether evidence is rated as clear seems more likely to reflect internal disclosure norms and the quality of governance within the firm, rather than being determined solely by the strength or positivity of the language used.

Outcome (Model)	Predictor	β	OR	p	Sig.
Promise	DSW	0.9767	2.656	<.001	***
Evidence Status	DSW	-0.4667	0.627	< .001	***
ESG Type (S vs E)	DSW	1.4258	4.161	<.001	***
ESG Type (G vs E)	DSW	-1.4917	0.225	<.001	***
Evidence Quality	ESW	0.0199	1.020	.856	n.s.

Note1. *** p < .001, ** p < .01, * p < .05; n.s. = not significant.
Note2. DSW = data_sentiment_weighted; ESW = evidence_sentiment_weighted.

Table 5: Regression Analysis Results for the Overall Dataset (Gemini 2.5 Pro)

Outcome (Model)	Industry	β	OR	p	Sig.
Promise	Semiconductor	0.9130	2.492	<.001	***
	Financial	0.8997	2.459	<.001	***
	Computer Periph.	1.0974	2.996	<.001	***
Evidence Status	Semiconductor	-0.4667	0.627	.0044	**
	Financial	-0.5140	0.598	.0012	**
	Computer Periph.	-0.5140	0.831	.2974	n.s.
ESG Type (S vs E)	Semiconductor	1.4474	4.252	<.001	***
	Financial	1.7474	5.740	<.001	***
	Computer Periph.	0.9635	2.621	<.001	***
ESG Type (G vs E)	Semiconductor	0.290	0.290	<.001	***
	Financial	-1.7016	0.182	<.001	***
	Computer Periph.	-1.2228	0.294	<.001	***
Evidence Quality	Semiconductor	0.0666	1.069	.7862	n.s.
	Financial	-0.2255	0.798	.2494	n.s.
	Computer Periph.	0.3082	1.361	.1553	n.s.

Note. *** p < .001, ** p < .01, * p < .05; n.s. = not significant.

Table 6: Cross-Industry Regression Analysis Results (Gemini 2.5 Pro)

Outcome (Model)	Predictor	β	OR	p	Sig.
Promise	DSW	0.3626	1.437	<.001	***
Evidence Status	DSW	-0.0341	0.967	.464	n.s.
ESG Type (S vs E)	DSW	1.5173	4.560	<.001	***
ESG Type (G vs E)	DSW	-0.5849	0.557	<.001	***
Evidence Quality	ESW	0.1399	1.150	.014	*

Note1. *** p < .001, ** p < .01, * p < .05; n.s. = not significant.
Note2. DSW = data_sentiment_weighted; ESW = evidence_sentiment_weighted.

Table 7: Regression Analysis Results for the Overall Dataset (FinBERT-ESG)

Outcome (Model)	Industry	β	OR	p	Sig.
Promise	Semiconductor	0.2247	1.252	<.001	***
	Financial	0.4116	1.509	<.001	***
	Computer Periph.	0.4430	1.557	<.001	***
Evidence Status	Semiconductor	0.1431	1.154	.0857	n.s.
	Financial	-0.1297	0.878	.0902	n.s.
	Computer Periph.	0.0539	1.055	.5844	n.s.
ESG Type (S vs E)	Semiconductor	1.9210	6.828	<.001	***
	Financial	1.2526	3.499	<.001	***
	Computer Periph.	1.5610	4.764	<.001	***
ESG Type (G vs E)	Semiconductor	-0.7512	0.472	<.001	***
	Financial	-0.5037	0.604	<.001	***
	Computer Periph.	-0.4071	0.666	<.001	***
Evidence Quality	Semiconductor	-0.0814	0.922	.5021	n.s.
	Financial	0.1859	1.204	.0385	*
	Computer Periph.	0.1690	1.184	.1292	n.s.

Note. *** p < .001, ** p < .01, * p < .05; n.s. = not significant.

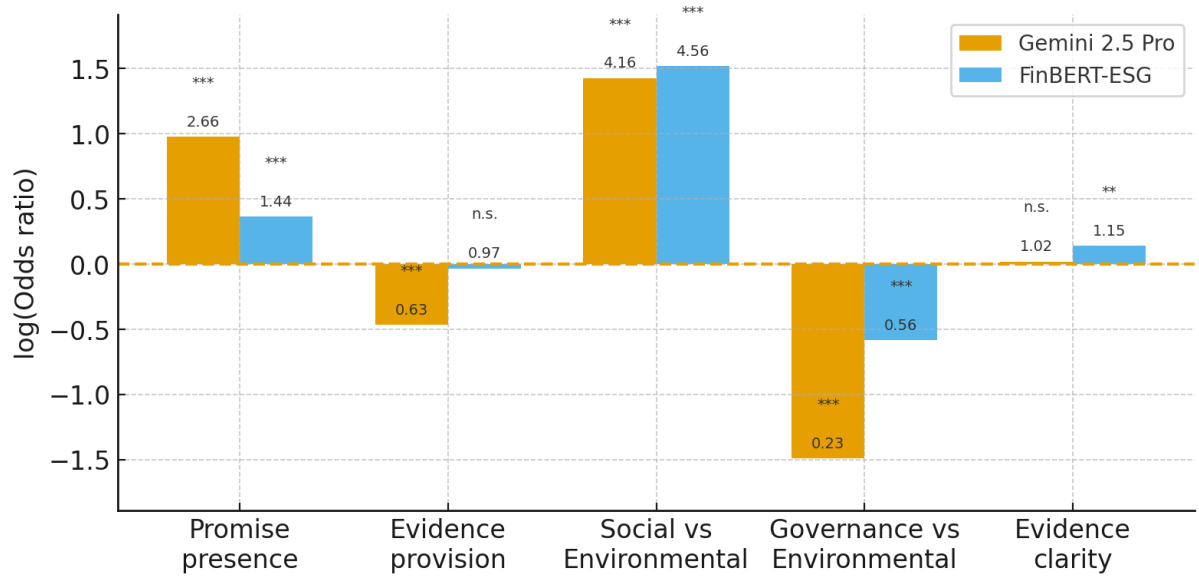
Table 8: Cross-Industry Regression Analysis Results (FinBERT-ESG)

4.5 Summary

Synthesizing the empirical findings reported in Tables 5 to 8, we can answer the four research questions as follows. Figure 2 provides a visual summary of these effects by plotting the log-odds ratios for both sentiment measures across all outcomes.

1. Sentiment and promise presence. Across the full sample and the industry-specific subsamples, and regardless of whether we rely on Gemini or FinBERT-ESG, sentiment scores exhibit a stable and statistically significant positive association with the presence of sustainability promises. In Figure 2, this is reflected in the clearly positive and significant bars for promise presence.
2. Sentiment and evidence provision. While the Gemini model and some industries reveal a negative association, this pattern is difficult to replicate consistently across models and industries, indicating that the impact of sentiment on evidence provision is not stable. Correspondingly, the bars for evidence provision in Figure 2 lie close to zero and differ in sign across models.
3. Sentiment and ESG-type distribution. The two sentiment models and all three industries show highly consistent directions and significance levels in the S vs. E and G vs. E models: positive sentiment is concentrated in social-dimension commitments, whereas governance-dimension commitments more often appear in text with weaker or more neutral sentiment. Figure 2 makes this pattern apparent through the strongly positive bars for Social vs. Environmental and the strongly negative bars for Governance vs. Environmental.
4. Evidence sentiment and clarity. Apart from a small positive effect in the FinBERT-ESG models for the full sample and one industry, the overall link between evidence sentiment and evidence clarity is weak. This suggests that clarity is more plausibly driven by institutional and governance factors than by sentiment alone. In Figure 2, the evidence-clarity bars are close to zero, with only a modest positive effect for FinBERT-ESG.

In summary, the analysis demonstrates that the tone of the language is strongly and consistently related to whether firms make promises and which ESG dimension those commitments fall into, but its association with whether evidence is provided and how clear that evidence is remains limited and unstable. Together with the regression tables, Figure 2 offers a concise overview of these patterns and provides a quantitative foundation and methodological reference for subsequent research on greenwashing, report quality, and automated text evaluation in the ESG domain.



Note. *** p < .001, ** p < .01, * p < .05; n.s. = not significant.

Figure2: Sentiment effects on all outcomes

Source：This study

V. CONCLUSION

This study develops a framework that combines large language models with dual-model sentiment analysis to examine whether the tone of sustainability disclosures is linked to firms' sustainability promise behavior. Using ESG reports from major Taiwanese listed companies, we construct the SentiPromiseESG dataset with 15,345 sentence-level annotations and derive weighted sentiment scores from Gemini 2.5 Pro and FinBERT-ESG. Logistic regression models are then used to evaluate how sentiment relates to the presence of promises, the provision of evidence, ESG-type classifications, and the clarity of supporting evidence.

The empirical results reveal several consistent patterns. First, sentiment scores show a stable positive association with the likelihood that a sustainability promise is made, regardless of model choice or industry. Second, the link between sentiment and evidence provision is less consistent: only Gemini shows a negative relationship in some industries, and this pattern does not hold across FinBERT-ESG or the broader sample. Third, sentiment clearly shapes the distribution of ESG promise types—more positive tone is more common in social-related commitments, while governance-related statements tend to appear in more neutral or less positive language. Fourth, sentiment in the evidence text has limited influence on evidence clarity, with only modest effects detected under FinBERT-ESG in the full sample and specific industries.

Overall, the findings suggest that tone plays an important role in signaling whether a firm chooses to make a sustainability commitment and which ESG dimension it highlights. In contrast, sentiment provides little insight into whether evidence is offered or whether that evidence is clearly articulated. Tone appears more reflective of communication style than of the underlying credibility of the commitments.

This study contributes to the literature in three ways. It introduces a scalable annotation pipeline that integrates LLM-based labeling with sentiment scoring from two distinct models. It provides the first large-scale empirical analysis linking sentiment tone to sustainability promise behavior and ESG narrative patterns. It also releases the SentiPromiseESG dataset, which offers a new resource for research on greenwashing, sustainability communication, and automated text evaluation.

Several limitations remain. Automated labeling may introduce noise, and sentiment outputs from different models can vary systematically. The dataset also focuses on Taiwanese firms in three major industries, which may limit generalizability. Future work may incorporate human validation, expand to international and multilingual ESG reports, integrate retrieval-augmented methods to improve evidence extraction, and develop longitudinal tools to track the fulfillment of sustainability promises.

In summary, this study provides a structured and reproducible approach for analyzing how sentiment relates to sustainability commitments and reporting practices. The results highlight the role of tone in shaping ESG narratives and point to new directions for evaluating disclosure quality and identifying potential greenwashing.

ACKNOWLEDGEMENTS

This research was supported by the Industrial Technology Research Institute (ITRI) and National Taipei University (NTPU), Taiwan, under grants NTPU-114A513E01 and NTPU-113A513E01; the National Science and Technology Council (NSTC), Taiwan, under grant NSTC 114-2425-H-305-003-; and National Taipei University (NTPU) under grant 114-NTPU_ORDA-F-004.

REFERENCES

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Atak, A. (2024). Beyond polarity: How ESG sentiment influences idiosyncratic volatility in the Turkish stock market. *Borsa Istanbul Review*, *24*, 10-21. https://doi.org/https://doi.org/10.1016/j.bir.2024.11.003

Barbeito-Caamaño, A., & Chalmeta, R. (2020). Using big data to evaluate corporate social responsibility and sustainable development practices. *Corporate Social Responsibility and Environmental Management*, *27*(6), 2831-2848. https://doi.org/https://doi.org/10.1002/csr.2006

Birti, M., Maurino, A., & Osborne, F. (2025). Optimizing large language models for esg activity detection in financial texts. Proceedings of the 6th ACM International Conference on AI in Finance,

Delmas, M. A., & Burbano, V. C. (2011). The Drivers of Greenwashing. *California Management Review*, *54*(1), 64-87. https://doi.org/10.1525/cmr.2011.54.1.64

He, L.-Y., & Wang, L. (2025). Can artificial intelligence curb greenwashing? Firm-level evidence based on large language model. *Energy Economics*, *152*, 108954. https://doi.org/https://doi.org/10.1016/j.eneco.2025.108954

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*, *40*(2), 806-841. https://doi.org/https://doi.org/10.1111/1911-3846.12832

Kim, S., Shin, Y., Park, S., Joel, S., Kim, S. T., & Oh, J. H. (2023). *Detecting Greenwashing in Sustainability Disclosures: A Prediction Model for KOSPI 200 Enterprises using ESG-BERT* 2023 IEEE International Conference on Big Data (BigData), https://doi.ieeecomputersociety.org/10.1109/BigData59044.2023.10386798

KPMG. (2022). *KPMG survey of sustainability reporting 2022. KPMG International.*

Liu, B. (2022). *Sentiment analysis and opinion mining*. Springer Nature.

LMArena. (2025). *LMArena leaderboard [Online resource]*. https://lmarena.ai/leaderboard

Lublóy, Á., Keresztúri, J. L., & Berlinger, E. (2025). Quantifying firm-level greenwashing: A systematic literature review. *Journal of Environmental Management*, *373*, 123399. https://doi.org/https://doi.org/10.1016/j.jenvman.2024.123399

Lyon, T., & Montgomery, A. (2015). The Means and End of Greenwash. *Organization & Environment*, *28*. https://doi.org/10.1177/1086026615575332

Martín-Domingo, L., Fernandez, J. B., Efthymiou, M., & Ali, M. I. (2025). Extracting airline emission KPIs from sustainability reports using large language models (LLMs). *Transportation Research Interdisciplinary Perspectives*, *33*, 101599. https://doi.org/https://doi.org/10.1016/j.trip.2025.101599

Ong, K., Mao, R., Xing, F., Satapathy, R., Sulaeman, J., Cambria, E., & Mengaldo, G. (2025). *ESGSenticNet: A Neurosymbolic Knowledge Base for Corporate Sustainability Analysis*. https://doi.org/10.48550/arXiv.2501.15720

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. https://doi.org/10.1016/j.frl.2024.104979

Seki, Y., Shu, H., Lhuissier, A., Lee, H., Kang, J., Day, M.-Y., & Chen, C.-C. (2024). ML-Promise: A Multilingual Dataset for Corporate Promise Verification. *arXiv preprint arXiv:2411.04473*.

Sun, Z., Satapathy, R., Guo, D., Li, B., Liu, X., Zhang, Y., Tan, C.-A., Filho, R. S., & Goh, R. S. M. (2024). *Information Extraction: Unstructured to Structured for ESG Reports* 2024 IEEE International Conference on Data Mining Workshops (ICDMW),

https://doi.ieeecomputersociety.org/10.1109/ICDMW65004.2024.00068

Turk, N., Khan, E., & Kosseim, L. (2025). CLaC at SemEval-2025 Task 6: A Multi-Architecture Approach for Corporate Environmental Promise Verification. *arXiv preprint arXiv:2505.23538*.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817-838. https://doi.org/10.2307/1912934

Xu, C., Miao, Y., Xiao, Y., & Lin, C. (2025). DeepGreen: Effective LLM-Driven Green-washing Monitoring System Designed for Empirical Testing--Evidence from China. *arXiv preprint arXiv:2504.07733*.