Multilingual Promise Verification in ESG Reports with Large Language Model Performance Evaluation

Wei-Chen Huang
Bachelor of Statistics
National Taipei University
New Taipei City, Taiwan
wesley@gmail.com

Hsin-Ting Lu
Graduate Institute of
Information Management
National Taipei University
New Taipei City, Taiwan
hsintinglubob@gmail.com

Wen-Ze Chen
Graduate Institute of
Information Management
National Taipei University
New Taipei City, Taiwan
50712andy@gmail.com

Min-Yuh Day*
Graduate Institute of
Information Management
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

Abstract

Corporate sustainability reports often contain vague or unverifiable statements, increasing the risk of greenwashing. As global expectations for the credibility of ESG disclosures continue to rise, developing automated systems capable of verifying corporate sustainability commitments has become an important research direction. However, current analytical approaches still face limitations in multilingual ESG promise verification, particularly in non-English language contexts.

This study investigates the performance of a large language model (GPT-5) in cross-lingual ESG promise verification tasks by evaluating corporate reports in Chinese, Japanese, and English, with the goal of establishing a multilingual evaluation benchmark. Four core subtasks are examined, including promise identification, evidence status assessment, evidence quality evaluation, and verification timeline prediction. Multiple prompting strategies—from zero-shot to few-shot learning, including Chain-of-Thought reasoning—are systematically compared to analyze the effectiveness of different design choices.

Results show that few-shot prompting generally yields more stable verification performance, while evidence quality evaluation remains the most challenging task across languages. Theoretically, this study proposes a cross-lingual prompting framework that clarifies how task complexity and annotation imbalance influence LLM reasoning performance in ESG verification. Practically, the findings provide actionable implications for regulators, investors, and corporate decision-makers by supporting the deployment of AI-based monitoring systems to enhance disclosure credibility, strengthen governance resilience, and enable more informed sustainable finance decisions.

Keywords: PromiseEval, Multilingual Dataset, Promise Verification, Greenwashing, Large Language Model

1 Introduction

This study aims to develop an advanced framework leveraging Large Language Model (LLM) to automatically verify whether corporate ESG reports contain explicit promises and to evaluate their credibility. Environmental, Social, and Governance (ESG) reporting has become a cornerstone of corporate accountability, with stakeholders increasingly relying on sustainability disclosures to inform investment decisions and assess corporate responsibility. However, the proliferation of ESG reporting has been accompanied by a concerning rise in greenwashing practices, where corporations overstate their environmental and social commitments while obscuring less favorable activities (Delmas & Burbano, 2011; Lyon & Montgomery, 2015). This phenomenon not only misleads stakeholders but also weakens the credibility of sustainability reporting.

Recent research emphasizes that greenwashing is both common and difficult to measure, as textual claims often lack clear evidence or measurable outcomes (Testa et al., 2018; Wang et al., 2025). To address this issue, computational approaches have been developed to automatically detect sustainability-related commitments and assess their validity. The PromiseEval shared task (Chen et al., 2025) introduced the first multilingual

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 303–313

benchmark for corporate promise verification, defining four subtasks:

1. **Promise Identification (PI):** determine whether a segment expresses promising contents.
2. **Supporting Evidence Assessment:** assess whether promises contain concrete evidence.
3. **Clarity of the Promise–Evidence Pair (CPEP):** evaluate the clarity and relevance of evidence in relation to the promise.
4. **Timing for Verification (TV):** indicate when a promise should be revisited for verification (e.g., within_two_years, two_to_five_years, more_than_5_years, others).

Building upon this foundation, the ML-Promise dataset (Seki et al., 2024) expanded multilingual coverage to five languages and incorporated retrieval-augmented generation techniques, demonstrating the feasibility of cross-lingual promise verification.

Despite progress in ESG analysis, significant analytical gaps persist. A systematic review by Lublóy et al. (2025) underscores the fragmented nature of current greenwashing quantification methods, a problem reflected in several key areas, like cultural and linguistic disparities, lack of integrated pipelines, insufficient verification baseline. Our study directly addresses these gaps by pioneering a multilingual framework that leverages Large Language Models (LLMs).

This study addresses these persistent analytical gaps, particularly concerning Chinese and Japanese reports, through three key contributions: (1) examining the feasibility of promise verification across multiple languages, (2) establishing baseline methods using state-of-the-art large language models for comparative analysis, and (3) providing methodological foundations for automated multilingual ESG verification systems.

2 Literature Review

2.1 ESG Reporting and Greenwashing

Greenwashing has been widely studied in sustainability communication research. Delmas and Burbano (2011) provide a conceptual framework for understanding the drivers of greenwashing, while Lyon and Montgomery (2015) emphasize its prevalence and regulatory implications. Empirical studies confirm that sustainability reports often contain misleading or unverifiable claims (Testa et al., 2018), reinforcing the need for computational tools. More recently, Wang, Gao, Wang et al. (2025) developed a greenwashing index using deep learning, providing quantitative evidence of discrepancies between corporate claims and substantiating evidence.

2.2 Computational Approaches to Detect Greenwashing

Recent advances in text mining and natural language processing (NLP) have been applied to analyze corporate sustainability disclosures and detect potentially misleading claims. For example, Wang Wang et al., (2025) proposed automated greenwashing indices derived from textual features of corporate reports, demonstrating how linguistic signals can indicate discrepancies between promises and actual practices. Beyond domain-specific applications, shared tasks such as SemEval-2022 Task 8 on Multilingual News Article Similarity (Chen et al., 2022) illustrates how NLP benchmarks can evaluate semantic consistency across texts in multiple languages. These computational approaches highlight the potential of AI-driven methods for large-scale monitoring of sustainability communication and for identifying unverifiable or vague ESG-related claims.

2.3 Corporate Promise Verification Tasks

The PromiseEval shared task, introduced at SemEval-2025 (Chen et al., 2025), formally established promise verification as a natural language processing (NLP) challenge. It defined tasks that align closely with the detection of vague or unverifiable claims in sustainability disclosures, focusing not only on promises but also on supporting evidence, clarity, and timeline. Its design highlights the complexity of promise verification and its close relationship to greenwashing detection

2.4 Multilingual NLP Datasets and Benchmarks

Multilingual benchmarks such as ML-Promise (Seki et al., 2024) have extended promise verification to multiple languages, addressing the gap in non-English corporate reporting. Other multilingual resources in NLP, such as XNLI (Conneau et al., 2018), show the value of multilingual evaluation, but ML-Promise is the first domain-specific dataset focused on corporate promises.

2.5 Annotation Quality and Inter-Annotator Agreement

Annotation quality is critical for promise verification tasks. Artstein and Poesio (2008) emphasize the role of inter-annotator agreement (IAA) metrics such as Cohen's Kappa and Krippendorff's Alpha to ensure annotation reliability. Both PromiseEval (Chen et al., 2025)and ML-Promise (Seki et al., 2024) adopted these metrics, reporting substantial agreement levels ($\kappa > 0.6$), which supports the validity of the datasets and subsequent analyses.

3 Experimental Setup

3.1 System Architecture

This study investigates multilingual promise evaluation using large language models as the foundational architecture. Promise evaluation encompasses the identification and verification of commitments within textual content, representing a critical component for assessing corporate statements in Environmental, Social, and Governance (ESG) reporting. Our research examines three linguistically diverse languages (Chinese, Japanese, and English), evaluating model performance across four distinct subtasks: Promise Identification (PI), Evidence Status Assessment (ESA), Evidence Quality Evaluation (EQE), and Verification Timeline Prediction (VTP).

To systematically assess model capabilities, we implement five prompting strategies: zero-shot, one-shot, three-shot, and five-shot learning, as well as an additional five-shot variant enhanced with Chain-of-Thought (CoT) prompting. Referencing the study by (Wei et al., 2022), we believe that Chain-of-Thought (CoT) requires sufficient examples to guide reasoning. Therefore, our analysis focuses on the 5-shot setting, as this configuration is not only our best-performing one, but also because the 5 examples provide sufficient context to allow us to isolate the effect of explicit reasoning. Within this specific setup, we systematically evaluate the marginal benefits of CoT. Building on this setup, the comprehensive evaluation framework enables controlled comparison of how demonstration quantity and reasoning instructions influence classification performance across different languages and verification tasks.(Figure 1 illustrates the overall research framework.)

3.2 Dataset

The study uses the PromiseEval dataset (Seki et al., 2024), which provides multilingual samples annotated for ESG-related promise verification. For each of the three languages (Chinese, Japanese, and English), the dataset is divided into 400 training samples and 400 test samples.

- Promise Status: classification of whether a concrete or organizational-level commitment is present.
- Evidence Status: detection of whether verifiable supporting evidence is provided.
- Evidence Quality: evaluation of evidence clarity (Clear, Not Clear, Misleading, N/A).
- Verification Timeline: identification of the expected timeline of promise fulfillment (Already, Within 2 years, Between 2–5 years, More than 5 years, N/A).

During the preliminary stage, samples were uniformly formatted for input to GPT-5. For few-shot conditions, demonstration examples were randomly sampled from the training set to prevent test data leakage and to simulate realistic evaluation scenarios.

In addition, to provide a clearer understanding of dataset composition, we analyzed the label distributions across the three languages. Table 1 and Table 2 present the distributions of the Chinese, Japanese, and English subsets. Although each training and test set contains the same number of samples (400 each), the label proportions across subtasks remain imbalanced, such as differences between positive and negative samples. These distributional characteristics may influence model classification performance.

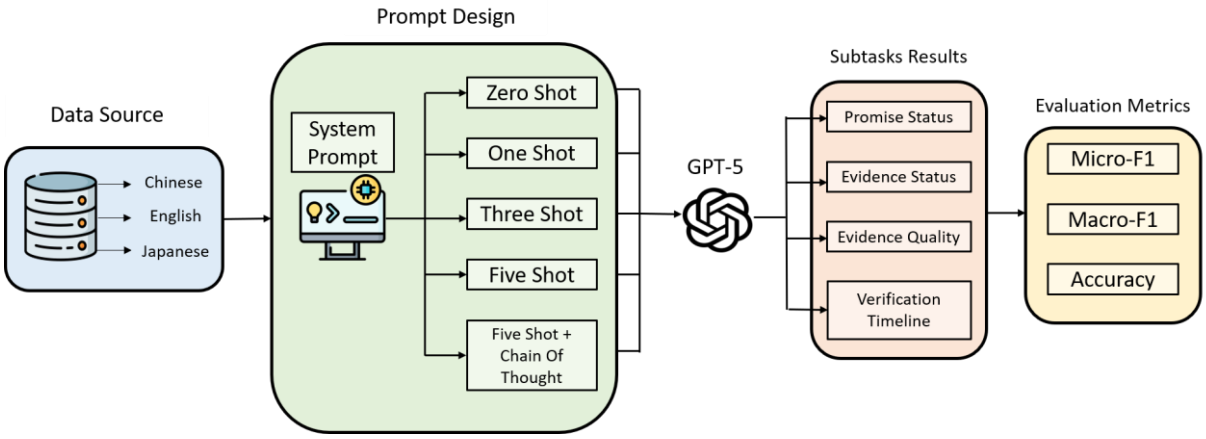

Figure 1: Proposed research workflow for ESG promise verification

Task	Label	Chinese	Japanese	English
Promise Status	Yes	146 (36.50%)	356 (89.00%)	313 (78.25%)
	No	254 (63.50%)	44 (11.00%)	87 (21.75%)
Evidence Status	Yes	78 (19.50%)	279 (69.75%)	221 (55.25%)
	No	322 (80.50)	77 (19.25%)	179 (44.75%)
	N/A	-	44(11.00%)	-
Evidence Quality	Clear	50(12.50%)	161(40.25%)	132(33.00%)
	Not Clear	16 (4.00%)	106 (26.50%)	85 (21.25%)
	Misleading	1 (0.25%)	12 (3.00%0	4 (1.00%)
	N/A	333 (83.25%)	121 (30.25%)	179 (44.75%)
Verification Timeline	Already	-	282 (70.50%)	155 (38.75%)
	Within 2 years	55 (13.75%)	18 (4.50%)	36 (9.00%)
	Between 2 to 5 years	7 (1.75%)	22 (5.50%)	75 (18.75%)
	More than 5 years	29 (7.25%)	34 (8.50%)	47 (11.75%)
	N/A	309 (77.25%)	44 (11.00%)	87 (21.75%)

Table 1: Label distribution of the PromiseEval training datasets (Chinese, Japanese, and English)

Task	Label	Chinese	Japanese	English
Promise Status	Yes	237 (48.47%)	372 (93.00%)	273 (68.25%)
	No	252 (51.53%)	28 (7.00%)	127 (31.75%)
Evidence Status	Yes	148 (30.27%)	232 (58.00%)	206 (51.50%)
	No	341 (69.73%)	140 (35.00%)	194 (48.50%)
	N/A	-	28 (7.00%)	-
Evidence Quality	Clear	73 (14.93%)	142 (35.50%)	134 (33.50%)
	Not Clear	46 (9.41%)	84 (21.00%)	71 (17.75%)
	Misleading	-	6 (1.50%)	1 (0.25%)
	N/A	370 (75.66%)	168 (42.00%)	194 (48.50%)
Verification Timeline	Already	-	295 (73.75%)	143 (35.75%)
	Within 2 years	101 (20.65%)	19 (4.75%)	36 (9.00%)
	Between 2 to 5 years	11 (2.25%)	17 (4.25%)	50 (12.50%)
	More than 5 years	39 (7.98%)	41 (10.25%)	44 (11.00%)
	N/A	338 (69.12%)	28 (7.00%)	127 (31.75%)

Table 2: Label distribution of the PromiseEval test datasets (Chinese, Japanese, and English)

3.3 Model and Strategies

We adopted GPT-5 as the unified Large Language Model (LLM) architecture across all languages and subtasks, focusing on evaluating the impact of prompt-based inference on classification performance. We designed five distinct prompting strategies.

First, regarding Prompting Strategies, we evaluated the following five settings:

- 0-shot: Consisted only of the task definition and system instructions in the prompt.
- 1-shot: The prompt was supplemented with one demonstration example.
- 3-shot: The prompt was supplemented with three demonstration examples.
- 5-shot: The prompt was supplemented with five demonstration examples.
- 5-shot + CoT (Chain-of-Thought): The prompt was supplemented with five demonstrations, along with an additional Chain-of-Thought instruction to encourage step-by-step logical reasoning before the final answer. However, the model was strictly required to output only the final structured label.

Second, concerning the Demonstration Source and Sampling for Few-Shot learning, all demonstration examples were selected from the training subset of the PromiseEval dataset to strictly prevent test data leakage and simulate realistic In-Context Learning evaluation scenarios. Given the significant class imbalance in our dataset (particularly for the Evidence Quality and Verification Timeline tasks), we employed a Stratified Random Sampling mechanism to select demonstrations. This ensured that the class label distribution in each Few-shot prompt (e.g., 'Yes'/'No' for Promise Status) maintained an approximate balance relative to the overall training set. This method aims to provide the LLM with a representative and stable context, thereby mitigating the class bias that pure random sampling might introduce.

This design enables a controlled comparison of how the number of demonstrations (from 0 to 5) and reasoning instructions (CoT) affect classification performance across languages and subtasks.

3.4 Evaluation Metrics

Model predictions on the test sets were compared against gold-standard annotations. Performance was measured using:

- Accuracy: Calculates the proportion of correctly predicted samples over the total number of samples, reflecting overall correctness at the instance level.
- Micro-F1: Aggregates true positives, false positives, and false negatives across all classes, reflecting overall predictive accuracy.
- Macro-F1: Computes F1-scores for each class independently, then averages them, ensuring fair evaluation of minority classes.

Together, these metrics provide a comprehensive assessment of both global accuracy and class-level robustness across multilingual promise evaluation tasks.

4 Experiment Results and Analysis

This section presents the performance evaluation of the GPT-5 model across the PromiseEval subtasks and provides an interpretative analysis within the context of the SemEval-2025 Task 6 shared task results. Our analysis covers performance metrics across three languages (Chinese, Japanese, and English) and five distinct prompting strategies, aiming to establish robust benchmarks for multilingual promise verification.

4.1 Overall Performance Analysis

Our results confirm that few-shot prompting consistently outperformed the zero-shot baseline, aligning with the principles of effective in-context learning. Table 3 shows that the 5-shot configuration is optimal, yielding the highest mean Accuracy (71.12%) and Macro-F1 (51.92%) across all tasks and languages. Conversely, the incorporation of Chain-of-Thought (CoT) reasoning led to a marginal and statistically non-significant decrease in aggregate performance (Accuracy 70.58%; Macro-F1 51.04%). However, a consistent downward trend was observed across multiple subtasks, suggesting that explicit reasoning did not consistently benefit pattern-based classification. A more detailed analysis of this phenomenon is provided in Section 4.4.

Strategy	Accuracy	Macro-F1	Micro-F1
0 shot	69.54%	47.66%	69.54%
1 shot	70.46%	49.95%	70.46%
3 shot	70.81%	51.37%	70.81%
5 shot	**71.12%**	**51.92%**	**71.12%**
5 shot_COT	70.58%	51.04%	70.58%

Table 3: Overall Performance Across All Tasks and Languages (Mean Values).

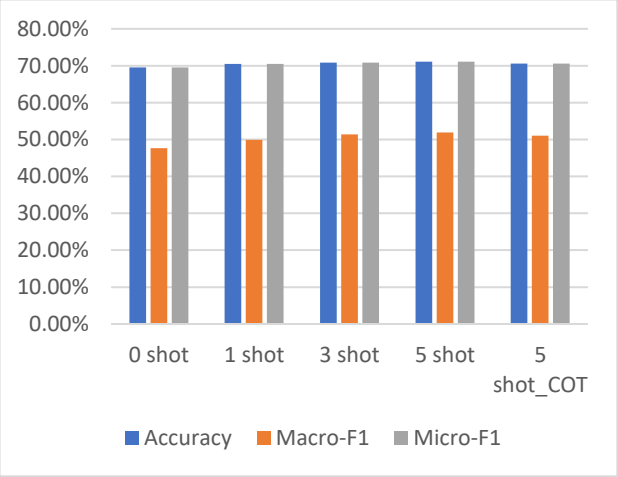

Figure 2: Overall Performance Across All Tasks and Languages (Mean Values).

4.2 Task-Specific Performance

4.2.1 Promise Status Identification

Promise Status Identification emerged as the most tractable subtask among those evaluated. Table 4 shows high performance across all languages, with the zero-shot setting in Japanese reaching 92.25% accuracy.

However, this high accuracy is a misleading artifact of severe class imbalance, where Table 1 shows the "Yes" class constitutes 89.8% of the Japanese dataset. This imbalance allows the model to achieve an inflated score by simply defaulting to the majority prediction. Consequently, the accuracy metric fails to penalize the model for its poor performance on the minority "No" class, a well-documented issue known as the Accuracy Paradox.

In contrast, the Macro F1 score provides a more robust evaluation by mitigating this bias. It achieves this by calculating the F1 score (which balances Precision and Recall) for each class independently before computing an unweighted

average, thus giving equal importance to both majority and minority classes. Under this more reliable metric, the 5-shot setting emerges as the top performer for Japanese with a Macro F1 score of 73.12%, outperforming the zero-shot setting's score of 71.64%. This trend, where multi-shot configurations yield superior Macro F1 scores, holds true across all languages, as Table 4 shows, confirming their ability to provide a more faithful assessment of true classification capabilities.

Language	Strategy	Accuracy	Macro-F1	Micro-F1
Chinese	0 shot	87.50%	85.39%	87.50%
	1 shot	91.00%	89.78%	91.00%
	3 shot	**91.50%**	**90.39%**	**91.50%**
	5 shot	91.00%	89.78%	91.00%
	5 shot_COT	91.00%	89.74%	91.00%
Japanese	**0 shot**	**92.25%**	71.64%	**92.25%**
	1 shot	89.00%	69.44%	89.00%
	3 shot	80.75%	63.01%	80.75%
	5 shot	91.75%	**73.12%**	91.75%
	5 shot_COT	89.50%	68.73%	89.50%
English	0 shot	75.50%	69.78%	75.50%
	1 shot	75.00%	68.83%	75.00%
	3 shot	77.25%	71.23%	77.25%
	5 shot	**78.00%**	**72.26%**	**78.00%**
	5 shot_COT	75.75%	71.84%	75.75%

Table 4: Promise Status Identification Performance by Language and Strategy.

Language	Strategy	Accuracy	Macro-F1	Micro-F1
Chinese	0 shot	82.25%	55.98%	82.25%
	1 shot	87.00%	73.79%	87.00%
	3 shot	87.75%	75.81%	87.75%
	5 shot	**88.00%**	**76.14%**	**88.00%**
	5 shot_COT	84.75%	69.89%	84.75%
Japanese	**0 shot**	**69.75%**	43.80%	**69.75%**
	1 shot	69.25%	44.60%	69.25%
	3 shot	69.00%	**45.43%**	69.00%
	5 shot	69.25%	44.94%	69.25%
	5 shot_COT	68.50%	44.61%	68.50%
English	0 shot	73.50%	72.80%	73.50%
	1 shot	74.75%	74.33%	74.75%
	3 shot	**75.25%**	**75.01%**	**75.25%**
	5 shot	72.75%	72.22%	72.75%
	5 shot_COT	73.50%	72.56%	73.50%

Table 5: Evidence Status Assessment Performance by Language and Strategy.

4.2.2 Evidence Status Assessment

Assessing Actionable Evidence requires relational reasoning and is substantially more complex than PI. Table 5 shows Chinese performance scaling with context, improving from 82.25% in the zero-shot setting to a peak of 88.00% in the 5-shot setting. In contrast, Table 5 also shows Japanese performance plateauing at 69.75% in the zero-shot setting and failing to improve with demonstrations. This stagnation suggests a high sensitivity to linguistic nuance that general LLM prompts struggle to capture. This observation is supported by the SemEval findings (Chen et al., 2025), where the WC Team achieved strong performance in Japanese evidence identification by utilizing the language-specific Tohoku-BERT model, suggesting that capturing language-specific writing styles is critical for evidence evaluation.

4.2.3 Evidence Quality Evaluation

Table 6 presents Evidence Quality Evaluation as the most challenging subtask, yielding the lowest and most variable Macro-F1 scores. Table 6 further shows the highest Macro-F1 in Chinese (40.98%) under the 5-shot + CoT strategy. This result confirms that explicit reasoning provides a marginal benefit in this fine-grained judgment task. The inferential difficulty of EQE is directly related to the assessment of misalignment. Table 1 indicates that misleading cases are rare in the full dataset—1 in Chinese and 23 in Japanese—yet they pose a significant risk and often involve superficial evidence or the linking of unrelated past data to future policies.

4.2.4 Verification Timeline Prediction

Verification Timeline Prediction yielded highly language-dependent outcomes. Table 7 indicates Chinese performance peaking at 83.00% Accuracy in the 5-shot setting, aligning with the distribution in which Chinese samples skew toward short-term verification. By contrast, Table 7 reports that few-shot learning did not improve English performance, with the zero-shot baseline remaining highest at 49.25%. Table 1 documents a large share of English samples labeled “Other” (245), suggesting indefinite timelines or non-temporal constraints that are underrepresented in the demonstrations. The task is further complicated by large corporations balancing short-term verification with long-term goals extending beyond five years, as Table 4 highlights.

4.3 Best-Case Performance by Language

Aggregating the optimal configurations reveals a marked cross-lingual disparity. Table 8 reports Chinese with the highest average Accuracy at 85.12%, substantially exceeding Japanese at 68.94% and English at 63.62%. This apparent Chinese advantage chiefly reflects dataset characteristics and severe class imbalance: SemEval analyses (Chen et al., 2025) note that the Chinese subset—owing to annotation methodology—contains a much lower proportion of positive samples. Table 1 documents this pattern in Actionable Evidence, where Chinese samples are predominantly labeled “No” (832 in the full dataset). As a result, Accuracy and Macro-F1 diverge widely for Chinese—Table 6 shows Evidence Quality at 78.00% Accuracy versus 40.98% Macro-F1—underscoring the need to treat Macro-F1 as the primary, less biased metric for fair cross-lingual comparison.

Language	Strategy	Accuracy	Macro-F1	Micro-F1
Chinese	0 shot	76.25%	39.35%	76.25%
	1 shot	77.50%	**41.01%**	77.50%
	3 shot	77.00%	39.71%	77.00%
	5 shot	77.50%	39.97%	77.50%
	5 shot_COT	**78.00%**	40.98%	**78.00%**
Japanese	0 shot	32.75%	21.16%	32.75%
	1 shot	37.25%	23.17%	37.25%
	3 shot	**38.50%**	**23.49%**	**38.50%**
	5 shot	36.25%	23.20%	36.25%
	5 shot_COT	36.50%	22.86%	36.50%
English	0 shot	44.25%	32.81%	44.25%
	1 shot	43.75%	32.16%	43.75%
	3 shot	**52.00%**	**37.15%**	**52.00%**
	5 shot	46.50%	33.39%	46.50%
	5 shot_COT	50.25%	36.40%	50.25%

Table 6: Evidence Quality Evaluation Performance by Language and Strategy.

Language	Strategy	Accuracy	Macro-F1	Micro-F1
Chinese	0 shot	76.75%	30.27%	76.75%
	1 shot	79.00%	35.83%	79.00%
	3 shot	80.75%	45.40%	80.75%
	5 shot	**83.00%**	**48.36%**	**83.00%**
	5 shot_COT	81.00%	43.22%	81.00%
Japanese	0 shot	74.50%	28.90%	74.50%
	1 shot	**75.25%**	27.76%	**75.25%**
	3 shot	74.50%	31.71%	74.50%
	5 shot	73.25%	30.64%	73.25%
	5 shot_COT	70.25%	**32.22%**	70.25%
English	**0 shot**	**49.25%**	**20.04%**	**49.25%**
	1 shot	46.75%	18.71%	46.75%
	3 shot	45.50%	18.10%	45.50%
	5 shot	46.25%	18.98%	46.25%
	5 shot_COT	48.00%	19.45%	48.00%

Table 7: Verification Timeline Prediction Performance by Language and Strategy.

Task	Chinese	Japanese	English
Promise Status	91.50%	**92.25%**	78.00%
Evidence Status	**88.00%**	69.75%	75.25%
Evidence Quality	**78.00%**	38.50%	52.00%
Verification Timeline	**83.00%**	75.25%	49.25%
Average	**85.12%**	68.94%	63.62%

Table 8: Best Performance by Task and Language (Highest Accuracy Configuration).

Task	With COT(Accuracy)	Without COT(Accuracy)	With COT(Macro-F1)	Without COT(Macro-F1)
Promise Status	85.42%	**86.92%**	76.77%	**78.39%**
Evidence Status	66.42%	**67.50%**	31.63%	**32.66%**
Evidence Quality	75.58%	**76.67%**	62.35%	**64.44%**
Verification Timeline	**54.92%**	53.42%	**33.41%**	32.19%

Table 9: Effectiveness of CoT Reasoning (Accuracy, %)

4.4 Impact of Chain-of-Thought Reasoning

Table 9 shows that the utility of CoT reasoning is highly task-dependent: when averaged across all languages and tasks, CoT yields a small but consistent aggregate decline—Accuracy decreases by 0.54 pp and Macro-F1 decreases by 0.88 pp. This confirms that CoT introduces unproductive processing overhead for tasks driven by direct semantic pattern matching. A phenomenon consistent with recent findings that step-by-step reasoning can actively degrade model accuracy in tasks resembling human overthinking scenarios (Liu et al., 2024).

However, CoT proved selectively effective, providing a measurable benefit in the Evidence Quality Evaluation subtask. This utility is maximized in inferentially complex scenarios demanding explicit, structured reasoning—such as assessing the likelihood of greenwashing. Therefore, CoT should be reserved for nuanced alignment tasks where multi-step judgment is required, rather than being applied as a default strategy for generalized classification.

5 Conclusion

This study aims to establish a multilingual evaluation framework for ESG promise verification using Large Language Models (LLMs) and to assess model performance across four verification subtasks in Chinese, Japanese, and English sustainability reports.

Our findings indicate that few-shot prompting, particularly the 5-shot configuration, provides more stable and reliable classification outcomes than zero-shot prompting, while the relative task difficulty differs significantly. Promise Identification is comparatively more tractable, whereas Evidence Quality Evaluation requires more complex contextual reasoning and remains the most challenging. Chain-of-Thought reasoning is not universally beneficial but demonstrates selective improvements in nuanced inference tasks.

In terms of academic contribution, this study provides a systematic benchmark for multilingual ESG promise verification and clarifies how task complexity, linguistic variation, and annotation imbalance jointly influence model reasoning behavior. It also enriches understanding of prompt design effects in cross-lingual sustainability contexts.

Regarding managerial implications, our results offer actionable guidance for ESG governance stakeholders. Organizations seeking to reduce greenwashing risks may adopt AI-driven verification mechanisms to enhance sustainability disclosure transparency and consistency, while regulatory bodies and investors can leverage these tools to improve oversight, credibility assessment, and accountability in sustainable finance.

Future work may incorporate retrieval-augmented techniques or domain-specific model adaptation to improve evidence relevance, extend evaluation to additional languages and industries, and develop explainable reasoning outputs to support real-world audits and compliance processes in sustainability reporting.

Acknowledgments

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

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555-596. https://doi.org/10.1162/coli.07-034-R2

Chen, C.-C., Seki, Y., Shu, H., Lhuissier, A., Kang, J., Lee, H., Day, M.-Y., & Takamura, H. (2025, July). SemEval-2025 Task 6: Multinational, Multilingual, Multi-Industry Promise Verification. In S. Rosenthal, A. Rosá, D. Ghosh, & M. Zampieri, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)* Vienna, Austria.

Chen, X., Zeynali, A., Camargo, C., Flöck, F., Gaffney, D., Grabowicz, P., Hale, S. A., Jurgens, D., & Samory, M. (2022, July). SemEval-2022 Task 8: Multilingual news article similarity. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, & S. Ratan, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)* Seattle, United States.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018, oct nov). XNLI: Evaluating Cross-lingual Sentence Representations. In E. Riloff, D. Chiang, J. Hockenmaier, & J. i. Tsujii, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* Brussels, Belgium.

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, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

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. P., & Montgomery, A. W. (2015). The Means and End of Greenwash. *Organization & Environment*, *28*(2), 223-249. https://doi.org/10.1177/1086026615575332

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

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*.

Testa, F., Boiral, O., & Iraldo, F. (2018). Internalization of Environmental Practices and Institutional Complexity: Can Stakeholders Pressures Encourage Greenwashing? *Journal of Business Ethics*, *147*(2), 287-307. https://doi.org/10.1007/s10551-015-2960-2

Wang, X., Gao, X., & Sun, M. (2025). Construction and analysis of corporate greenwashing index: a deep learning approach. EPJ Data Science, 14(1), 44. https://doi.org/10.1140/epjds/s13688-025-00562-w

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024). Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. arXiv:2410.21333. https://doi.org/10.48550/arXiv.2410.21333