SemEval-2025 Task 6: Multinational, Multilingual, Multi-Industry Promise Verification

Chung-Chi Chen,[1] Yohei Seki,[2] Hakusen Shu,[3] Anaïs Lhuissier,[4] Juyeon Kang,[4] Hanwool Lee,[5] Min-Yuh Day,[6] Hiroya Takamura[1]

[1]AIST, Japan

[2]Institute of Library, Information, and Media Science, University of Tsukuba, Japan

[3]College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba, Japan

[4]3DS Outscale, France [5]Shinhan Securities Co., Korea

[6]Graduate Institute of Information Management, National Taipei University, Taiwan

Abstract

While extensive research exists on misinformation and disinformation, there is limited focus on future-oriented commitments, such as corporate Environmental, Social, and Governance (ESG) promises, which are often difficult to verify yet significantly impact public trust and market stability. To address this gap, we introduce the task of promise verification, leveraging natural language processing (NLP) techniques to automatically detect ESG commitments, identify supporting evidence, and evaluate the consistency between promises and evidence, while also inferring potential verification time points. This paper presents the dataset used in SemEval-2025 PromiseEval, outlines participant solutions, and discusses key findings. The goal is to enhance transparency in corporate discourse, strengthen investor trust, and support regulators in monitoring the fulfillment of corporate commitments.

1 Introduction

In an era characterized by rapid information dissemination and increasing reliance on public statements from influential figures–such as corporate executives and political leaders–the balance between freedom of speech and ethical responsibility has become a critical societal concern. While freedom of speech empowers individuals to express opinions and make commitments, it also raises complex challenges when these statements impact public trust, financial decisions, or social stability. Only a small number of political leaders, such as the president, may be tracked manually (Waller and Morieson, 2025).[1] However, when this extends to legislators or even corporate-level individuals, the number of targets increases significantly, making automated tracking a necessary and inevitable approach.

Although there has been extensive discussion on disinformation (Alam et al., 2022; Vykopal et al., 2024; Pan et al., 2024) and misinformation (Qazvinian et al., 2011; Wu et al., 2022; Yang et al., 2023; Ma et al., 2024), there is relatively little focus on statements related to the future. While some analyses of forward-looking statements (Chen and Takamura, 2024; Lin et al., 2024) address future events, promises represent visions for the future that are more abstract in nature and difficult to verify. For example, a forward-looking statement might forecast the next quarter's earnings per share (EPS), which can be verified once the quarter concludes. However, a promise may be less specific, such as "the company will continue to strive to reduce carbon emissions in the coming years." Such a promise can create a positive impact for the company but also poses regulatory challenges.

Public figures often make promises that shape societal expectations and influence critical decisions. These promises are typically based on the information available at the time. However, when such commitments go unfulfilled, the consequences extend beyond personal accountability to affect broader public trust and market stability. The challenge lies in discerning whether these unfulfilled promises result from unforeseen circumstances, representing legitimate changes in strategy, or if they were knowingly misleading from the outset. In the financial field, the regulatory framework governing corporate disclosures has long established sophisticated guidelines for forward-looking statements, yet remains underdeveloped in addressing ESG commitments and promises.

This regulatory asymmetry persists despite ESG commitments increasingly resembling financial forward-looking statements in their market impact. More companies now disclose climate transition plans with quantified milestones, and some institutional investors use ESG forward-looking metrics in capital allocation decisions. Yet, unlike financial forward-looking statements, ESG-related

[1]Example: https://www.politifact.com/truth-o-meter/promises/

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2461–2471

Task	Label	English	French	Chinese	Japanese	Korean
Promise Identification	Yes	755	764	464	898	155
	No	245	236	635	102	45
Actionable Evidence	Yes	549	646	267	621	146
	No	451	354	832	277	47
Clarity of Promise-Evidence Pair	Clear	327	440	147	365	128
	Not Clear	212	197	75	233	7
	Misleading	10	9	1	23	0
	Other	451	354	876	-	-
Timing for Verification	Within 2 years	76	64	187	48	65
	2-5 years	150	166	26	55	12
	Longer than 5 years	105	95	81	104	25
	Other	245	236	805	0	41
	Already	424	439	-	691	-

Table 1: Dataset Statistics

promises face significant legal and ethical risks. Some climate-related statements have been alleged as cases of "greenwashing."

As an early step in assessing the integrity and fulfillment of the company's ESG promises, we introduce a new task: promise verification, which considers multinational, multilingual, and multi-industry aspects. We propose leveraging natural language processing (NLP) techniques to automatically detect ESG commitments made by companies, identify supporting evidence, and evaluate the alignment between the stated commitments and the corresponding evidence. Furthermore, this approach aims to infer or detect potential time points at which these commitments can be verified. This fine-grained analysis ensures transparency and accountability in corporate ESG discourse. Through multilingual and multi-industry scalability, the system is designed to operate across geopolitical boundaries and sector-specific nuances, making it adaptable to diverse regulatory environments and cultural expectations. This research aims to foster a more informed and accountable public sphere.

In this paper, we present the details of the dataset used in SemEval-2025 PromiseEval, the solutions from task participants, and the findings of this round's shared task. We believe that this dataset and the proposed systems can not only assist regulators but also help companies prepare reports that are more trustworthy to investors, as well as support companies in detecting any incompleteness in the information disclosed through natural language.

2 Task and Dataset

2.1 Task Definition

We propose four core tasks–Promise Identification, Actionable Evidence, Clarity of the Promise-Evidence Pair, and Timing for Verification–to serve as a foundational framework for evaluating corporate Environmental, Social, and Governance commitments. Table 1 lists the label design for each task.

1. **Promise Identification**: This is a boolean label (*Yes/No*) used to determine whether a promise is present in the statement. A promise can take the form of a declaration outlining a company principle (e.g., diversity and inclusion), a commitment (e.g., reducing plastic waste, enhancing health & safety), or a strategic initiative (e.g., protocol descriptions, establishing partnerships with associations and institutes) that aligns with ESG (Environmental, Social, and Governance) criteria. It is crucial to distinguish between substantive promises and superficial statements, as companies may make broad claims without tangible backing, which is particularly important when assessing the risk of greenwashing. If there is no foundational statement describing a principle or commitment, it should not be classified as a promise.

2. **Actionable Evidence**: This boolean label (*Yes/No*) evaluates whether concrete evidence exists that demonstrates the company is actively working towards fulfilling its promise. Valid evidence includes specific examples, implemented measures, quantitative data, reports, or third-party audits that support the promise. Documentation such as tables, pie charts, or statistical reports serve as quantified evidence, enhancing the credibility of a textual core promise. The absence of such supporting material raises concerns about the company's transparency and accountability.

3. **Clarity of the Promise-Evidence Pair**: This

Industry	English	French	Chinese	Japanese	Korean
Energy	✓	✓	✓	✓	✓
Finance	✓	✓			✓
Luxury	✓	✓			
Semiconductor			✓		✓
Technology			✓		✓
Biomedical			✓		✓
Automotive				✓	✓
Trading				✓	

Table 2: Industry-based statistics

Country	English	French	Chinese	Japanese	Korean
UK	✓				
USA	✓				
Jordan	✓				
South Africa	✓				
Switzerland	✓				
Canada	✓	✓			
France	✓	✓			
Luxembourg		✓			
Taiwan			✓		
Japan				✓	
South Korea					✓

Table 3: Country-based statistics

criterion is evaluated using three possible labels (*Clear/Not Clear/Misleading*) and focuses on the strength of the connection between the promise and its supporting evidence. A *Clear* label indicates that the provided evidence is both specific and sufficient to substantiate the promise. A *Not Clear* label reflects ambiguity or insufficient detail, making it difficult to confirm the company's commitment. The *Misleading* label is applied when the evidence appears intentionally deceptive or when it misrepresents the actual fulfillment of the promise. Both the quantity and quality of evidence are critical in this assessment.

4. **Timing for Verification**: Adhering to Morgan Stanley Capital International (MSCI) guidelines[2] and prior research (Tseng et al., 2023; Chen et al., 2024), this label outlines when stakeholders should re-evaluate the promise to verify its fulfillment. The following time frames are used: *within 2 years*, *2-5 years*, *longer than 5 years*, and *other*. The *other* category is used when a promise has already been verified, is ongoing without a definitive timeline, or when no specific future verification is required. This labeling ensures that stakeholders can monitor ESG-related actions within an appropriate timeframe, promoting accountability and long-term impact assessment.

2.2 Data Analysis

Table 1 presents the distribution of labels across the dataset. The distribution of labels for the Promise Identification task indicates that most samples in English, French, Japanese, and Korean were identified as containing a promise. In contrast, the Chinese data exhibited a significantly lower proportion of promises, with the majority labeled as "No."

For the Actionable Evidence task, Korean and French samples showed a higher presence of actionable evidence, suggesting that commitments in these languages are often accompanied by concrete supporting details. English and Japanese demonstrated moderate levels of actionable evidence. The Chinese data again reflected a lower presence, primarily due to differences in the annotation approach. While the Chinese subset marks each page of the ESG report, other languages focus on specific sections where promises are more likely to appear. This discrepancy stems from annotation coverage rather than report structure.

In terms of the clarity between promises and their corresponding evidence, the Korean data exhibited exceptionally high clarity, indicating strong alignment between commitments and supporting details. English, French, and Japanese showed moderate clarity levels, while Chinese maintained a relatively high level despite lower rates in previous tasks. The "Misleading" label was minimal across all languages.

Regarding the timing of commitments, Korean data strongly favored short-term verifications. Chinese data also leaned toward shorter timelines. French data showed a preference for long-term commitments. A substantial portion of English data was classified as "Other," suggesting either indefinite timelines or commitments not bound by explicit temporal constraints.

This analysis reveals distinct linguistic and cultural patterns across the dataset. Korean data is characterized by high clarity and a short-term orientation, while French commitments tend to imply long-term planning. These findings underscore the importance of considering language-specific characteristics in promise analysis and may reflect broader cultural norms.

We provide industry-based and country-based statistics in Tables 2 and 3. These tables reveal

[2] https://www.msci.com/sustainability-and-climate-methodologies

Task	Label	Market Cap		
		High	Medium	Low
Promise Identification	Yes	52.78%	40.12%	23.23%
	No	47.22%	59.88%	76.77%
Actionable Evidence	Yes	25.14%	29.01%	16.54%
	No	74.86%	70.99%	83.46%
Clarity of Promise-Evidence Pair	Clear	66.29%	59.14%	80.49%
	Not Clear	32.58%	40.86%	19.51%
	Misleading	1.12%	0.00%	0.00%
Timing for Verification	Within 2 years	54.64%	78.21%	78.79%
	2-5 years	4.92%	12.82%	21.21%
	Longer than 5 years	40.44%	8.97%	0.00%

Table 4: Distribution across different market capitalization – Chinese

that the dataset covers a wide range of sectors and geographic regions, highlighting the diversity and representativeness of the collected promises.

From an industry perspective, the Energy sector is selected across all languages. The Finance and Luxury industries are primarily covered in English and French datasets, consistent with the prominence of European and North American firms in these sectors. Meanwhile, high-technology industries such as Semiconductors, Technology, and Biomedical are particularly prominent in the Chinese dataset, aligning with the strategic economic focus in Taiwan. Automotive and Trading industries are notably present in the Japanese dataset, reflecting Japan's global leadership in these areas.

Country-wise, the English dataset aggregates statements from a diverse range of countries including the United Kingdom, United States, Jordan, South Africa, Switzerland, and Canada, showing a broad international spread. French data mainly originates from France, Canada, and Luxembourg, reflecting the global Francophone economic landscape. Chinese, Japanese, and Korean datasets are primarily sourced from Taiwan, Japan, and South Korea respectively.

This wide coverage ensures that the dataset is not only multilingual but also multicultural and multisectoral, allowing researchers to study promise verification in a variety of economic, regulatory, and cultural contexts. Such diversity is crucial for building robust, generalizable models and for enabling fine-grained analyses that account for regional and sector-specific nuances in corporate ESG discourse.

2.3 Market Capitalization

Previous studies (Cormier and Magnan, 2003; Hahn and Kühnen, 2013) have indicated that the quality of sustainability and ESG reporting may be related to company size. Larger companies, measured by asset size, number of employees, and market capitalization, tend to produce higher-quality and more transparent ESG reports due to their abundant resources and greater external pressures from investors, governments, and media scrutiny. In contrast, smaller companies, which typically face limited resources and lower external pressure, may produce ESG reports that are less detailed and accurate. Therefore, when selecting companies for our analysis, we categorized firms within each industry into high, medium, and low market capitalization groups. Table 4 presents the statistical summary of the Chinese dataset. Analysis of the distribution reveals notable differences across company sizes. For the Promise Identification task, large companies had a higher number of positive identifications than medium-sized and small companies, indicating that larger firms are more likely to explicitly make ESG promises. In the Actionable Evidence task, the proportions were relatively similar across company sizes, suggesting that regardless of company size, firms demonstrated comparable levels of effort in providing concrete evidence to support their ESG promises.

Regarding the Clarity of the Promise-Evidence Pair, small companies demonstrated the highest clarity rate, followed by large and then medium-sized companies. This suggests that although smaller companies make fewer promises, the ones they do make tend to be more clearly supported by evidence, possibly due to more focused ESG initiatives or simpler reporting structures. For the Timing for Verification, small and medium-sized companies favored short-term verifications within two years, whereas large companies had a more balanced distribution between short-term verification and longer-term goals extending beyond five years. This pattern indicates that larger corporations often set long-term sustainability objectives, while smaller firms prefer immediate or near-term

Task	Label	Industry		
		Semiconductor	Energy	Biomedical
Promise Identification	Yes	48.90%	58.24%	18.11%
	No	51.10%	41.76%	81.89%
Actionable Evidence	Yes	28.61%	30.00%	14.17%
	No	71.39%	70.00%	85.83%
Clarity of Promise-Evidence Pair	Clear	60.00%	62.38%	88.57%
	Not Clear	40.00%	36.63%	11.43%
	Misleading	0.00%	0.99%	0.00%
Timing for Verification	Within 2 years	59.87%	70.27%	66.67%
	2-5 years	1.27%	14.41%	19.05%
	Longer than 5 years	38.85%	15.32%	14.29%

Table 5: Distribution across different industry – Chinese

demonstrable achievements.

Overall, the findings suggest that company size plays a critical role in shaping ESG communication practices. Larger firms are more proactive in making promises and pursuing long-term strategies, while smaller firms emphasize clarity and short-term execution in their commitments.

2.4 Industry-Wise Analysis

To further understand sector-specific differences in ESG communication, we conducted an industry-wise analysis based on the Chinese dataset, focusing on Semiconductor, Energy, and Biomedical sectors. The statistics are provided in Table 5.

In the Promise Identification task, the Energy sector exhibited the highest proportion of promises, followed by the Semiconductor sector, and finally the Biomedical sector. This suggests that Energy companies are more proactive in articulating their ESG commitments, likely due to increasing regulatory and societal pressures regarding environmental impact. In contrast, Biomedical companies demonstrated a notably lower rate of ESG promise articulation, possibly reflecting a more cautious or conservative disclosure strategy. For the Actionable Evidence task, both Semiconductor and Energy sectors showed moderate levels of actionable evidence, indicating that although promises are made, the provision of concrete supporting evidence remains limited. Biomedical companies further highlight the challenges in demonstrating tangible ESG progress in highly regulated and research-driven industries.

Regarding the Clarity of the Promise-Evidence Pair, the Biomedical sector stood out with the highest clarity, significantly surpassing both Semiconductor and Energy sectors. This result implies that although Biomedical companies make fewer promises, they tend to ensure a strong and clear linkage between their commitments and supporting documentation, possibly due to stricter compliance standards in the healthcare domain. In terms of Timing for Verification, all three sectors heavily favored short-term verification within two years, particularly the Energy sector. This trend suggests a focus on near-term accountability, driven by stakeholder demand for demonstrable ESG progress. Semiconductor companies showed a notable proportion of long-term commitments, reflecting the sector's need for longer innovation cycles and infrastructure development to achieve sustainability goals.

Overall, the industry-wise analysis reveals that sector-specific dynamics significantly influence how ESG promises are formulated, supported, and communicated. The Energy sector, under intense scrutiny, emphasizes frequent and near-term disclosures; the Semiconductor sector balances short- and long-term perspectives; while the Biomedical sector prioritizes clarity over quantity in its ESG messaging. These findings underline the importance of tailoring ESG verification systems to industry-specific characteristics to ensure fair and accurate assessments.

3 Participants and Methods

The SemEval-2025 Task 6 attracted various innovative approaches from participating teams, reflecting the diverse strategies used to tackle multilingual ESG promise verification. Ten teams shared their experimental results on the Kaggle leaderboard, and seven teams submitted a method description paper. Some teams participated only in the Promise and Evidence verification tasks, while others focused on specific languages. We summarize their methods in this section. Please refer to the paper for more details.

CSECU-DSG (Hossain and Chy, 2025) proposed a Bi-LSTM-based model, which leverages

Team Name	Method
CSECU-DSG (Hossain and Chy, 2025)	Bi-LSTM (LASER and Universal Embedding)
CSCU (Leesombatwathana et al., 2025)	GPT-4o, SVM, DistilBERT (Data Augmentation)
CYUT (Wu et al., 2025)	Llama-3.1 (Structured Prompt) & RAG
Oath (Khubaib et al., 2025)	DeBERTa (Data Augmentation)
QM-AI (Sun and Sobczak, 2025)	Ensemble BERT (Data Augmentation)
WC Team (Nishi and Takagi, 2025)	BERT (CamemBERT, Tohoku-BERT)
YNU-HPCC (deng et al., 2025)	BERT (R-Drop)

Table 6: Overview of methods.

Subtask	Participant	Overall	Promise	Evidence	Clarity	Timing
English	CSCU	0.678	0.760	0.779	0.648	0.526
	Baseline	0.677	0.760	0.787	0.639	0.519
	Oath	0.661	0.739	0.770	0.669	0.465
	CYUT	0.649	0.775	0.674	0.549	0.577
	CT	0.619	0.746	0.702	0.592	0.437
	QM-AI	0.519	0.823	0.786	0.218	0.247
	ConU	0.522	0.702	0.694	0.519	0.174
	YNU-HPCC	0.442	0.587	0.516	0.395	0.271
	WC Team	0.375	0.787	0.714	–	–
	CSECU-DSG	0.326	0.701	0.406	0.005	0.194
French	CYUT	0.677	0.822	0.753	0.593	0.542
	Baseline	0.661	0.764	0.762	0.615	0.503
	QM-AI	0.541	0.832	0.791	0.281	0.258
	WC Team	0.372	0.724	0.764	–	–
	CSECU-DSG	0.313	0.646	0.432	0.0003	0.173
Korean	Baseline	0.636	0.820	0.827	0.761	0.136
	QM-AI	0.549	0.835	0.760	0.592	0.007
	CSECU-DSG	0.066	0.152	0.110	–	–
Chinese	SemanticEval	0.561	0.504	0.604	0.610	0.526
	Baseline	0.360	0.580	0.503	0.434	0.526
	QM-AI	0.353	0.683	0.565	0.070	0.093
	CSECU-DSG	0.323	0.617	0.674	–	–
Japanese	Baseline	0.606	0.912	0.648	0.427	0.731
	QM-AI	0.531	0.925	0.667	0.251	0.281
	CSECU-DSG	0.492	0.896	0.445	0.0005	0.626
	WC Team	0.402	0.921	0.686	–	–

Table 7: Experimental results

both LASER and Universal Sentence Encoder embeddings to capture multilingual semantic features. By combining these embeddings and using Bi-LSTM to capture sequential patterns, their approach focuses on improving cross-lingual promise identification performance.

CSCU (Leesombatwathana et al., 2025) focused on data augmentation, particularly through Paraphrase Augmentation and Synthesis Augmentation generated by Gemini-2.0-Flash. They compared multiple classifiers, including GPT-4o (zero-shot and six-shot), Support Vector Machine (SVM) using Multilingual E5 embeddings, and fine-tuned DistilBERT. Their results highlight that synthetic data augmentation significantly boosts classification performance, especially for identifying promises and supporting evidence.

CYUT (Wu et al., 2025) adopted a Structured Prompting with Retrieval-Augmented Generation (RAG) approach, using Llama 3.1 to systematically evaluate promises. Their framework employs structured definitions, examples, and step-by-step reasoning, combined with retrieval of relevant context, to enhance ESG promise verification across multiple languages.

Oath (Khubaib et al., 2025) introduced a DeBERTa-based pipeline, enhanced with contrastive learning and data augmentation to handle class imbalance and subtle differences between promise-related and non-promise text. This approach significantly improved evidence classification and timeline verification performance.

QM-AI (Sun and Sobczak, 2025) proposed an ensemble BERT framework, combining four different BERT variants (original, augmented, translated, and mixed-language versions). This ensemble approach leverages both multilingual training and language-specific fine-tuning to improve robustness, particularly for non-English data.

WC Team (Nishi and Takagi, 2025) adopted a

language-specific BERT strategy, training individual BERT models for each language—BERT-base-uncased for English, CamemBERT for French, and Tohoku-BERT for Japanese. This monolingual fine-tuning approach emphasizes capturing each language's unique syntactic and semantic characteristics, achieving strong performance for high-resource languages.

Finally, YNU-HPCC (deng et al., 2025) introduced a BERT model regularized with R-Drop, a regularization technique that forces consistent predictions across different dropout applications. This method stabilizes the model's predictions in low-resource and noisy environments, particularly benefiting smaller language datasets within the task.

These methods collectively demonstrate the importance of data augmentation, multilingual embeddings, structured prompting, and regularization techniques when developing robust promise verification systems for ESG reports across diverse languages.

4 Evaluation Results

Each team was allowed to submit multiple entries. The public leaderboard reflected performance on 70% of the test set, while the remaining 30% was kept private by the organizers. The values presented in Table 7 were recalculated based on each team's highest-scoring entry on the private leaderboard, using the entire test set (both the public and private portions). Therefore, these recalculated results may differ slightly from the original public leaderboard results. Additionally, since some teams only participated in predicting specific columns, scores for columns without submissions are marked as "–." Ten teams shared their experimental results on the Kaggle leaderboard, and seven teams submitted a method description paper. Some teams participated only in the Promise and Evidence verification tasks, while others focused on specific languages.

As shown in Table 7, Team CSCU ranked first for English data, achieving high overall scores on all tasks using synthetic data augmentation and GPT-4o (final submission) which outperformed the fine-tuned DistilBERT model and SVM (Leesombatwathana et al., 2025). For French, Team CYUT placed 1st using LLaMA 3.1:70b, enhanced by structured prompting along with RAG and CoT strategies (Wu et al., 2025). Team QM-AI placed first for both Korean and Japanese data using an ensemble of fine-tuned BERT-base models with data augmentation (Sun and Sobczak, 2025). For Chinese, Team SemanticEval placed first (no paper was submitted).

The participants experimented diverse approaches, ranging from transformer architecture to LLM-based framework, revealing important insights about effective multilingual NLP strategies. Most results demonstrate the effectiveness of models designed with multilingual capabilities and the potential of advanced LLM techniques such as RAG, CoT, GoT Structured Prompting when properly implemented. Additionally, Team WC achieved encouraging results with a monolingual approach using separate models for each language, while Team Oath experimented with different approaches for each subtask, showing promising scores compared to the multilabel classification approach.

The varying effectiveness of data augmentation and ensemble methods across languages and tasks highlights the need for language/task-specific strategies rather than one-size-fits-all approaches. For English and French, data augmentation showed minimal benefits contrary to the Japanese, Korean and Chinese datasets for which data augmentation using an LLM like GPT-4o provided meaningful improvements. The best-performing approaches combined targeted data augmentation, fine-tuning on large models like BERT, DeBERTa and XLM-RoBERTa and an LLM framework, which is obviously a viable alternative to traditional fine-tuning.

For Korean, the highest performance was achieved by Team QM-AI, which utilized an ensemble of multilingual BERT models directly trained on original texts without English translation (Sun and Sobczak, 2025). A notable challenge specific to the Korean dataset was that texts were provided as PDF pages rather than extracted snippets, requiring additional preprocessing to extract clean textual inputs. Although data augmentation using GPT-4o generally improved results for Korean, excessive synthetic augmentation negatively affected performance, as observed with Team CSECU-DSG (Hossain and Chy, 2025). This indicates the importance of carefully balancing data augmentation and preserving linguistic nuances specific to Korean.

For Japanese, three teams—QM-AI, CSECU-DSG, and WC Team—participated. Their approaches included:

	Promise						Evidence					
	Yes			No			Yes			No		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SemanticEval	0.33	0.18	0.23	0.70	0.84	0.77	0.33	0.18	0.23	0.70	0.84	0.77
CSECU-DSG	0.50	0.31	0.38	**0.74**	**0.87**	**0.80**	**0.50**	0.31	0.38	0.74	**0.87**	**0.80**
QM-AI	**0.79**	**0.50**	**0.61**	0.65	0.88	0.75	0.36	**0.64**	**0.46**	**0.77**	0.51	0.61
Ensemble (Vote)	0.48	0.30	0.37	0.74	0.86	0.79	0.48	0.30	0.37	0.74	0.86	0.79
Heuristic (No First)	0.70	0.59	0.64	0.67	0.77	0.71	0.52	0.23	0.32	0.73	0.91	0.81
Heuristic (Yes First)	0.69	0.62	0.66	0.68	0.74	0.71	0.36	0.66	0.47	0.77	0.50	0.60

Table 8: Fine-grained comparison of models in the Chinese dataset.

- QM-AI: BERT-based ensemble models with data augmentation and machine translation
- CSECU-DSG: Dual embedding models (Laser and Universal Sentence Encoder) combined with LSTM and MLP
- WC Team: Japanese-specific Tohoku BERT

From the results, QM-AI achieved the highest scores in the Promise Identification and Clarity of Promise-Evidence Pair subtasks, CSECU-DSG led in Timing Verification, and WC Team excelled in Actionable Evidence Identification. QM-AI's insights suggest that while data augmentation and machine translation were ineffective, their ensemble approach generalized well for certain subtasks. Conversely, universal sentence embeddings were particularly effective for timing verification, while a Japanese-specific approach performed well for evidence identification, likely due to language-specific writing styles. Notably, machine translation to English was less effective than in previous tasks (Chen et al., 2024), possibly because the Promise Verification task required deeper language-specific knowledge.

5 Discussion

5.1 Data Imbalance and Labeling Challenges

We observe that participants encountered data imbalance issues both between languages and across subtasks, as most experimented with multilingual and multilabel classification approaches. This is also reflected in the final scores as shown in Table 7: the Promise Identification subtask showed consistently high scores across languages, which also coincides with higher agreement among annotators. In contrast, other labels, particularly Clarity (Clarity of Promise-Evidence Pair) and Timing (Timing for Verification), posed greater challenges, likely due to higher subjectivity and lower agreement among annotators.

To address data imbalance, future work should go beyond collecting more balanced datasets and refining guidelines - given that ESG data preparation requires expert involvement and validation - by also exploring and leveraging LLM-based cross-lingual data augmentation techniques (Whitehouse et al., 2023) to enhance representation in low-resource languages.

5.2 Comparison of Models

Table 8 presents the performance of different systems for the Chinese subtask, including SemanticEval, CSECU-DSG, QM-AI, and an ensemble voting method. The metrics cover precision (P), recall (R), and F1-score (F1) for both *Yes* and *No* labels under each task. The main findings are summarized as follows:

- **QM-AI excels in identifying promises and evidence**: QM-AI achieves the highest F1 for the *Yes* class in both tasks, demonstrating its superior ability to detect substantive promises and concrete evidence.
- **CSECU-DSG is reliable for rejecting non-promises and missing evidence**: CSECU-DSG achieves the highest F1 for the *No* class, indicating strong performance in identifying irrelevant or unsupported statements.

The ensemble model, built by majority voting, does not significantly outperform the strongest single system (QM-AI). In particular, its F1 for the *Yes* class is lower than that of QM-AI, indicating that ensemble voting diluted the strength of QM-AI's positive predictions. While ensemble voting slightly improves the *No* class F1, this gain does not offset the drop in identifying positive cases.

This result highlights a fundamental limitation: simple voting ensembles are not effective when there is a large performance gap between models. The weaker models (e.g., SemanticEval) pull down the overall performance, especially in the crucial *Yes* class. This is problematic for ESG promise detection, where identifying actual promises and evidence is more critical than filtering out irrelevant content. The findings suggest that a better ensemble strategy would apply weighted voting, where

QM-AI contributes more to the *Yes* class decisions, while CSECU-DSG could contribute more to the *No* class. We further provide Heuristic results in Table 8. This adaptive weighting would better reflect each model's strengths, potentially leading to a more robust and interpretable final system.

In conclusion, QM-AI is the most effective standalone system, especially for detecting positive cases, which is crucial for identifying real ESG commitments and evidence. The current ensemble approach, however, fails to add value and may even hurt performance. Future work should explore weighted or task-specific ensemble strategies to better combine the complementary strengths of different models. The choice of method and the aspect to prioritize ultimately depends on the specific application scenario. For instance, if the goal is to verify the existence of promises, detecting the presence of a promise (the *Yes* class) becomes the primary objective. On the other hand, for regulatory agencies, the focus might shift toward identifying cases where no concrete evidence is provided (the *No* class), as such instances signal potential transparency issues or greenwashing risks. Therefore, the relative importance of detecting *Yes* versus *No* should align with the intended use case and stakeholder needs.

5.3 Case Studies – Misleading

Although misleading cases are relatively rare in the dataset, they represent significant risks to transparency and accountability. One common pattern involves presenting superficial evidence, such as emphasizing trust with suppliers to imply material quality without objective proof. In other instances, companies announce future policies alongside unrelated past data, creating an illusion of responsiveness. Ambiguous promises, like support for low- and moderate-income communities, sometimes cite general charity work rather than directly addressing the original ESG goal. Similarly, companies may highlight employee training without clearly linking it to the technical innovations previously mentioned. In each case, the misalignment between promise and evidence can mislead stakeholders, inflating perceptions of progress without substantive backing.

5.4 Future Directions

To further enhance the depth and effectiveness of the proposed direction, two additional dimensions can be incorporated: Risk of Greenwashing and Stakeholder Impact. These dimensions should also be aligned with the proposed four tasks to ensure a comprehensive and cohesive approach.

The Risk of Greenwashing expands the assessment by evaluating the likelihood that a company's ESG claims are misleading or lack substantive backing. By categorizing promises into *Low*, *Moderate*, or *High* risk based on the consistency between the stated commitment and supporting evidence, the clarity of communication, and the involvement of third-party verification, this criterion helps stakeholders critically appraise the authenticity of Environmental, Social, and Governance statements. A *High* risk rating signals vague promises with little or no verifiable data, while a *Low* risk reflects transparent and substantiated claims.

In parallel, the Stakeholder Impact dimension assesses the extent to which an ESG promise directly or indirectly benefits stakeholders. Classifications of *Direct*, *Indirect*, or *Minimal* impact allow for a nuanced understanding of the promise's reach and relevance. *Direct* impacts refer to immediate effects on employees, customers, or local communities (e.g., enhancing workplace safety), while *Indirect* impacts relate to broader societal or environmental outcomes (e.g., reducing carbon emissions). *Minimal* impact reflects limited or negligible stakeholder benefits.

Together, these extended dimensions complement the original four tasks, providing a more comprehensive evaluation of Environmental, Social, and Governance commitments. They enable stakeholders to not only verify the authenticity and clarity of corporate promises but also assess their broader societal implications and the risk of misleading information.

6 Conclusion

This paper presents SemEval-2025 Task 6, a multilingual, multi-industry shared task on verifying corporate ESG promises—a shift from misinformation research to future corporate commitments critical for public trust. We introduced a novel dataset in five languages annotated for promise identification, evidence detection, clarity, and timing. Results from diverse methods, including multilingual transformers and LLM-enhanced prompting, highlight challenges like data imbalance and linguistic variation. We propose expanding research to greenwashing risk and stakeholder impact, aiming to strengthen ESG transparency and accountability.

Acknowledgments

The work of Chung-Chi Chen and Hiroya Takamura was supported in part by the AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain." This work of Yohei Seki was partially supported by the JSPS the Grant-in-Aid for Scientific Research (B) (#23K28375) and by ROIS NII Open Collaborative Research 2024 (24S1204).

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chung-Chi Chen and Hiroya Takamura. 2024. Term-driven forward-looking claim synthesis in earnings calls. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15752–15760, Torino, Italia. ELRA and ICCL.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anais Lhuissier, Yohei Seki, Hanwool Lee, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multilingual ESG impact duration inference. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 219–227, Torino, Italia. Association for Computational Linguistics.

Denis Cormier and Michel Magnan. 2003. Environmental reporting management: A continental european perspective. *Journal of Accounting and Public Policy*, 22(1):43–62.

dehui deng, You Zhang, Jin Wang, Dan Xu, and Xuejie Zhang. 2025. Ynu-hpcc at semeval-2025 task 6: Using bert model with r-drop for promise verification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1895–1901, Vienna, Austria. Association for Computational Linguistics.

Rüdiger Hahn and Michael Kühnen. 2013. Determinants of sustainability reporting: A review of results, trends, theory, and opportunities in an expanding field of research. *Journal of Cleaner Production*, 59:5–21.

Tashin Hossain and Abu Nowshed Chy. 2025. Csecu-dsg at semeval-2025 task 6: Exploiting multilingual feature fusion-based approach for corporate promise verification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1842–1852, Vienna, Austria. Association for Computational Linguistics.

Muhammad Khubaib, Owais Aijaz, and Ayesha Enayat. 2025. Oath breakers at semeval-2025 task 06: Promiseeval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1711–1716, Vienna, Austria. Association for Computational Linguistics.

Kittiphat Leesombatwathana, Wisarut Tangtemjit, and Dittaya Wanvarie. 2025. Cscu at semeval-2025 task 6: Enhancing promise verification with paraphrase and synthesis augmentation: Effects on model performance. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1925–1937, Vienna, Austria. Association for Computational Linguistics.

Chin-Yi Lin, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Argument-based sentiment analysis on forward-looking statements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13804–13815, Bangkok, Thailand. Association for Computational Linguistics.

Weicheng Ma, Chunyuan Deng, Aram Moossavi, Lili Wang, Soroush Vosoughi, and Diyi Yang. 2024. Simulated misinformation susceptibility (SMISTS): Enhancing misinformation research with large language model simulations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2774–2788, Bangkok, Thailand. Association for Computational Linguistics.

Takumi Nishi and Nicole Miu Takagi. 2025. Wc team at semeval-2025 task 6: Promiseeval: Multinational, multilingual, multi-industry promise verification leveraging monolingual and multilingual bert models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1663–1669, Vienna, Austria. Association for Computational Linguistics.

Tsung-Hsuan Pan, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Enhancing society-undermining disinformation detection through fine-grained sentiment analysis pre-finetuning. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1371–1377, St. Julian's, Malta. Association for Computational Linguistics.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Zihang Sun and Filip Sobczak. 2025. Qm-ai at semeval-2025 task 6: an ensemble of bert models for promise identification in esg context. In *Proceedings of the*

19th International Workshop on Semantic Evaluation (SemEval-2025), pages 232–237, Vienna, Austria. Association for Computational Linguistics.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5412–5416.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. Disinformation capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14830–14847, Bangkok, Thailand. Association for Computational Linguistics.

Lisa Waller and Lucy Morieson. 2025. Election promise tracking: Extending the shelf life of democracy in digital journalism practice and scholarship. *Journalism Studies*, pages 1–17.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.

Shih-Hung Wu, Zhi-Hong Lin, and Ping-Hsuan Lee. 2025. Cyut at semeval-2025 task 6: Prompting with precision – esg analysis via structured prompts. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 497–504, Vienna, Austria. Association for Computational Linguistics.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Sin-han Yang, Chung-chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Entity-aware dual co-attention network for fake news detection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 106–113, Dubrovnik, Croatia. Association for Computational Linguistics.