ML-Promise: A Multilingual Dataset for Corporate Promise Verification

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

[1]Institute of Library, Information, and Media Science, University of Tsukuba, Japan,
[2]Graduate School of Comprehensive Human Sciences, University of Tsukuba, Japan,
[3]3DS Outscale, France, [4]Shinhan Securities Co., Korea,
[5]Graduate Institute of Information Management, National Taipei University, Taiwan
[6]AIST, Japan

Abstract

Promises made by politicians, corporate leaders, and public figures have a significant impact on public perception, trust, and institutional reputation. However, the complexity and volume of such commitments, coupled with difficulties in verifying their fulfillment, necessitate innovative methods for assessing their credibility. This paper introduces the concept of Promise Verification, a systematic approach involving steps such as promise identification, evidence assessment, and the evaluation of timing for verification. We propose the first multilingual dataset, ML-Promise, which includes English, French, Chinese, Japanese, and Korean, aimed at facilitating in-depth verification of promises, particularly in the context of Environmental, Social, and Governance (ESG) reports. Given the growing emphasis on corporate environmental contributions, this dataset addresses the challenge of evaluating corporate promises, especially in light of practices like greenwashing. Our findings also explore textual and image–based baselines, with promising results from retrieval-augmented generation (RAG) approaches.[1]

1 Introduction

In a world where promises shape perceptions and drive decisions, the integrity of commitments made by politicians, corporate leaders, and public figures must be scrutinized. These promises, ranging from environmental sustainability to social responsibility and governance ethics, significantly influence the general public's and stakeholders' trust, as well as government and corporate reputations. Yet, the complexity and abundance of such commitments, coupled with the challenge of verifying their fulfillment, create a pressing need for innovative approaches to assess their strength and verifiability. Recognizing the critical role of transparency and accountability in today's society, we propose a groundbreaking task: Promise Verification.

To perform promise verification, several steps are required, including (1) identifying the promise, (2) linking the promise with actionable evidence, (3) assessing the clarity of the promise–evidence pair, and (4) inferring the timing for verifying the promise. Evaluating the quality of ESG-related promises requires assessing the availability of evidence demonstrating a company's commitment to fulfilling them. The clarity of this evidence directly influences the perceived credibility of the promise. Therefore, a precise definition of evidence clarity is essential. For example, Santos, a gas company, claims it will achieve net-zero emissions by 2040. However, this claim has been challenged by a citizen group, arguing that it relies on unproven carbon capture and storage technologies.[2] In this case, the evidence supporting the company's promise can be classified as "not clear". Additionally, whether the author provides a clear timeline for verifying the promise is an important criterion. For instance, "we will achieve net zero carbon emissions within five years" is a stronger promise than "we will achieve net zero carbon emissions." Following this line of thought, this paper proposes the first multilingual dataset for in-depth promise verification, including Chinese, English, French, Japanese, and Korean.

In recent years, increasing emphasis has been placed on companies' environmental contributions, especially in addressing climate change, deforestation, and compliance with labor conditions and governance, when evaluating their investment value. In the evolving landscape of ESG (environmental, social, and governance) criteria, the ability to accurately assess a company's promises and adherence to its ESG promises has become paramount. However, unlike traditional financial statements, ESG

[1]Dataset: http://MLpromise.nlpfin.com/

[2]https://www.edo.org.au/2021/08/26/world-first-federal-court-case-over-santos-clean-energy-net-zero-claims/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20365–20378
November 4-9, 2025

reports still lack clear standards regarding corporate promises. This allows some companies to use misleading information to project an overly positive environmental image, a practice known as greenwashing. As Gorovaia and Makrominas (2024) point out, companies involved in environmental misconduct tend to produce longer, more positive, and more frequent reports. We hypothesize that such reports may lack substantive evidence, or the information presented may be irrelevant or ambiguous, leading to misinterpretation. To this end, the proposed dataset, ML-Promise, focuses on ESG reports released by corporations in five countries: the U.K., France, Taiwan, Japan, and Korea.

To provide a comprehensive benchmark for promise verification, ML-Promise comprises 3,010 annotated instances (2,010 for training and 1,000 for testing across five languages), with labels for Promise Identification, Actionable Evidence, Clarity of the Promise–Evidence Pair, and Timing for Verification. The dataset was curated from ESG reports of companies across diverse industries, ensuring linguistic and contextual variability.

Beyond text–based baselines, we also explore image–based approaches, recognizing that most ESG reports are published in PDF format. Our experiments incorporate retrieval-augmented generation (RAG) (Lewis et al., 2020) to enhance performance. The results indicate that RAG improves clarity assessment and timing prediction but exhibits language-dependent variations in promise identification and actionable evidence detection. Furthermore, our dataset reveals notable differences in ESG reporting styles across regions, underscoring the need for multilingual and multimodal analysis in promise verification.

2 Related Work

2.1 ESG Report Analysis

Recent studies have sought to improve the analysis of ESG or sustainability reports for estimating company values using contextual embedding approaches. For example, Gutierrez-Bustamante and Espinosa-Leal (2022) evaluated sustainability reports from publicly listed companies in Nordic countries using latent semantic analysis (LSA) and the global vectors for word representation (GloVe) model, enhancing document retrieval performance based on similarity. Garigliotti (2024) explored the integration of sustainable development goals (SDGs) into environmental impact assessments (EIAs) using a RAG framework powered by large language models (LLMs). Their work focused on two tasks: detecting SDG targets within EIA reports and identifying relevant textual evidence, specifically in European contexts. Hillebrand et al. (2023) introduced sustain.AI, a context-aware recommender system designed to analyze sustainability reports in response to increasing corporate social responsibility (CSR) regulations. The system, based on a BERT architecture, identified relevant sections of lengthy reports using global reporting initiative (GRI) indicators and demonstrated strong performance on datasets from German companies. We organized the ML-ESG shared-task series, which aimed to estimate how long the effects of certain events or actions taken by a company will last, impacting its ESG scores Chen et al. (2024a). However, their study did not establish a promise verification framework. Additionally, their dataset did not focus on ESG reports.

Previous studies have a few shortcomings. First, most of them focus solely on reports from one country. Second, none of them attempt to analyze corporate promises, despite the abundance of sustainability reports. Third, they primarily examine sustainability reports and social media rather than ESG reports. While sustainability reports outline goals and strategies, climate reports focus on climate-related actions, and annual reports may include ESG sections, they often lack a comprehensive overview. Company websites and social media platforms rarely provide exhaustive information. In contrast, ESG reports serve as formal documents dedicated to a company's ESG initiatives and, more importantly, their outcomes—whether the company has met its stated goals. As such, they provide the most reliable evidence for assessing corporate accountability.

To address these problems, our study extends these works by focusing on multilingual companies from both European and Asian regions, including Taiwan, the UK, France, Japan, and Korea. With the proposed new task, we aim to highlight the importance of anti-greenwashing by evaluating corporate promises in ESG reports. Recent fact-checking research has also focused on annotating evidential information (Chen et al., 2024b; Drchal et al., 2024). In conventional evidence retrieval for fact-checking, the common approach is to extract supporting or refuting evidence from a collection of documents using methods based on semantic similarity or coreference resolution. In contrast, our

task focuses on identifying supporting evidence for corporate promises within the same document, specifically documents published by the companies themselves. The primary goal is to clarify corporate stances in ESG-related communications. Furthermore, our approach addresses linguistic variations depending on whether the claim pertains to environmental, social, or governance issues. Thus, our task and corpus offer distinctive and noteworthy value beyond existing fact-checking studies. Building on these insights, we assess the clarity of evidence supporting ESG commitments to address greenwashing concerns. Additionally, our methodology incorporates visual elements to capture all possible evidence, enhancing the credibility of our findings.

2.2 Retrieval-Augmented Generation

RAG (Lewis et al., 2020) was introduced as a method to enhance LLMs by integrating external knowledge sources. This approach combines retrieval mechanisms with generative models, producing more accurate and contextually relevant outputs. Yu et al. (2024) highlight the advantages of RAG systems, particularly their ability to extract domain-specific information. Fan et al. (2024) discuss training strategies for RAG, including independent, sequential, and joint methods, which can be tailored to optimize retrieval and generation for specific domains. For Chinese language applications, Wang et al. (2024b) emphasize the importance of domain-specific corpora over general knowledge sources. Ardic et al. (2024) applied RAG to analyze sustainability reports from ten Turkish companies, focusing on ESG factors. Following the findings of previous studies, we also explore the RAG approach as a proof of concept and design it for the proposed tasks.

3 ML-Promise

3.1 Task Design

We collect ESG reports from five countries: the UK, France, Taiwan, Japan, and Korea. We chose three major industries per country, selecting three companies from each, resulting in ESG reports from nine companies per country. The annotators are native speakers of the target language or are familiar with the language at the work level. The task designs are as follows when given an instance[3] in the ESG reports. We provide some examples in Appendix A.

1. **Promise Identification (PI)**: This is a boolean label (Yes/No) based on whether a promise exists. A promise can be a statement, which states a company principle (e.g., diversity and inclusion), commitment (e.g., reducing plastic waste, improving health & safety) or strategy (e.g., protocol description, developing partnerships with associations and institutes) related to ESG criteria.
2. **Actionable Evidence (AE)**: This is a boolean label (Yes/No) based on whether the intended evidence for the company taking action towards fulfilling the promise exists. The evidence deemed the most relevant to prove the core promise is being kept, which includes simple examples, company measures, numbers, etc. Tables and charts included in the report count as quantitative evidence that supports a textual promise.
3. **Clarity of the Promise–Evidence Pair (CPEP)**: We designed three labels (Clear/Not Clear/Misleading) for this task, which should depend on the clarity of the given evidence in relation to the promise. The clarity is the assessment of the company's ability to back up their statement with enough clarity and precision. Note that clarity is defined by a combination of quantity and quality of evidence.
4. **Timing for Verification (TV)**: Following the MSCI guidelines and previous work (Tseng et al., 2023), we set timing labels (within 2 years/2–5 years/longer than 5 years/other) to indicate when readers/investors should return to verify the promise. This is the assessment of when we could possibly see the final results of a given ESG-related action and thus verify the statement. Here, "other" denotes the promise has already been verified or doesn't have a specific timing to verify it.

3.2 Industries and Companies

This study incorporates a cultural dimension by examining how ESG (Environmental, Social, and Governance) criteria and reporting practices vary across regions. While certain industries prioritize specific ESG aspects—such as environmental concerns in the Energy sector—and face unique regulatory challenges, all industries are ultimately subject

[3]We define the instance as the unit corresponding to the paragraph(s) containing the promise and the evidence(s).

Task	Label	English		French		Chinese		Japanese		Korean		Total	
		#	%	#	%	#	%	#	%	#	%	#	%
Promise Identification	Yes	169	84.5	161	80.5	80	40.2	149	74.9	155	77.5	714	71.5
	No	31	15.5	39	19.5	119	59.8	50	25.1	45	22.5	284	28.5
Actionable Evidence	Yes	122	61.6	141	71.6	40	20.1	99	66.4	146	75.6	548	58.5
	No	76	38.4	56	28.4	159	79.9	50	33.6	47	24.4	388	41.5
Clarity of Promise–Evidence Pair	Clear	56	53.3	77	56.6	22	64.7	60	61.2	128	94.8	343	67.5
	Not Clear	45	42.9	57	41.9	12	35.3	34	34.7	7	5.2	155	30.5
	Misleading	4	3.8	2	1.5	0	0.0	4	4.1	0	0.0	10	2.0
Timing for Verification	Within 2 years	3	1.9	19	12.4	30	37.5	11	7.3	65	45.5	128	18.8
	2–5 years	22	14.1	23	15.0	8	10.0	14	9.3	12	8.4	79	11.6
	Longer than 5 years	14	9.0	33	21.6	12	15.0	28	18.7	25	17.5	112	16.4
	Other	117	75.0	78	51.0	30	37.5	97	64.7	41	28.7	363	53.2

Table 1: Label distribution by language. (number of labels (#) and percentages (%))

to the same standards for clarity and compliance. Therefore, we evaluate different industries under uniform criteria while incorporating multiple layers of comparison, including country, industry, and company size. Industries were selected based on their significance in the participating countries and their frequent discussion in international ESG summits. This includes sectors like Energy and Finance/Economy. To enhance comparability, a third industry was selected. This industry reflects each country's economic identity—for example, Luxury for France.

To deepen the analysis, we examined companies of three different sizes and market shares within each industry. This approach allows us to assess how company size and market influence affect ESG compliance and greenwashing practices. Additionally, only recent ESG reports (from 2021 onward) were included to align with current ESG reporting regulations, ensuring the study's relevance. The selection of three companies per industry was based on varying market capitalizations to highlight differences in the writing styles of ESG promises and actionable evidence across companies with different market values.

For the Korean dataset, due to the limited availability of ESG-related textual materials from small companies, 29 major corporations were included, encompassing large conglomerates (Chaebols) such as Samsung, SK, Hyundai, LG, LOTTE, and Doosan, as well as leading venture companies like Kakao, Naver, and HYBE. This selection provides a more comprehensive representation of ESG reporting trends in Korea.

3.3 Inter-annotator Agreement by Task

The detailed annotation process is shown in Appendix B. Cohen's κ inter-annotator agreement (Cohen, 1960; McHugh, 2012) for each classification attribute is summarized as follows. Across tasks, κ ranges from 0.65–0.96 (PI), 0.71–0.88 (AE), 0.62–0.80 (CPEP), and 0.60–0.89 (TV). All values exceed 0.60—considered *substantial* agreement (Landis and Koch, 1977)—indicating reliable annotations.

3.4 Statistics

Finally, we obtained 3,010 instances, i.e., 600 for each language and 10 additional instances in the Chinese dataset. Table 1 presents the distribution of the proposed ML-Promise dataset. First, we observe that around 35-40% of the evidence is "not clear" in supporting the associated promises in four out of five languages. This highlights the necessity of the proposed task for evaluating the quality of the promise–evidence pairs from corporations. Furthermore, about 4% of instances contain (potentially) misleading evidence in the English and Japanese datasets. It is crucial for corporations to re-examine this evidence, and it is also essential for supervisory authorities to monitor these instances. Second, we noted that corporations in Taiwan and Korea tend to make more short-term promises (within 2 years), whereas corporations in the remaining countries tend to make longer-term promises. This finding shows the need for a multilingual comparison of ESG reports across different countries, as the narrative styles vary among them.

4 Data Analysis

This section provides a comprehensive synthesis and interpretation of the statistical data on corporate ESG commitments included in the ML-Promise dataset. The analysis goes beyond simple data presentation to offer a multifaceted perspective on multinational and multilingual ESG reporting patterns by integrating information from various tables.

Task		Chinese			French			English			Japanese			Korean		
		High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Promise Id.	Yes	52.78%	40.12%	23.23%	78.23%	79.30%	72.36%	72.16%	71.01%	82.42%	90.28%	89.12%	90.00%	83.44%	80.48%	77.78%
	No	47.22%	59.88%	76.77%	21.77%	20.70%	27.54%	27.84%	28.99%	17.58%	9.72%	10.88%	10.00%	16.56%	19.52%	22.22%
Act. Evid.	Yes	25.14%	29.01%	16.54%	65.62%	65.83%	61.79%	49.14%	48.12%	65.94%	78.77%	69.64%	57.04%	37.76%	41.39%	37.04%
	No	74.86%	70.99%	83.46%	34.38%	34.17%	38.21%	50.86%	51.88%	34.07%	21.23%	30.36%	42.96%	62.24%	58.61%	62.96%
Clarity	Clear	66.29%	59.14%	80.49%	56.25%	78.09%	69.74%	55.24%	51.81%	67.50%	67.97%	54.03%	50.00%	55.17%	54.17%	53.85%
	Not Clear	32.58%	40.86%	19.51%	41.82%	20.95%	28.94%	42.66%	45.18%	31.67%	29.69%	41.71%	44.81%	34.48%	41.67%	38.46%
	Misleading	1.12%	0.00%	0.00%	1.93%	0.96%	1.32%	2.10%	3.01%	0.83%	2.34%	4.27%	5.19%	10.34%	4.17%	7.69%
Timing	< 2 yrs	54.64%	78.21%	78.79%	6.46%	8.43%	10.11%	6.67%	11.02%	11.67%	6.46%	5.60%	4.44%	13.79%	7.14%	10.00%
	2-5 yrs	4.92%	12.82%	21.21%	19.36%	22.09%	23.60%	21.90%	21.63%	17.00%	4.92%	10.07%	4.44%	10.34%	7.14%	10.00%
	> 5 yrs	40.44%	8.97%	0.00%	13.71%	5.22%	17.98%	11.90%	18.78%	11.33%	12.62%	13.06%	10.37%	10.34%	4.76%	10.00%
	Other	–	–	–	60.48%	64.26%	48.31%	59.52%	48.57%	60.00%	76.00%	84.33%	80.74%	65.52%	80.95%	70.00%

Table 2: Distribution across different market capitalization.

Task		Chinese			French			English			Japanese			Korean		
		Sem.	Energy	Bio.	Finance	Energy	Luxury	Finance	Energy	Luxury	Auto	Energy	Trading	Sem.	IT	Holdings
Promise Id.	Yes	48.90%	58.24%	18.11%	68.14%	78.70%	83.49%	70.50%	80.12%	75.95%	92.71%	91.99%	84.73%	92.11%	87.88%	75.00%
	No	51.10%	41.76%	81.89%	31.86%	21.30%	16.51%	29.50%	19.88%	24.05%	7.29%	8.01%	15.27%	7.89%	12.12%	25.00%
Act. Evid.	Yes	28.61%	30.00%	14.17%	56.23%	68.21%	70.48%	48.25%	54.68%	62.34%	72.46%	62.58%	72.79%	44.74%	51.52%	41.67%
	No	71.39%	70.00%	85.83%	43.76%	31.79%	29.52%	51.75%	45.32%	37.66%	27.54%	37.42%	27.21%	55.26%	48.48%	58.33%
Clarity	Clear	60.00%	62.38%	88.57%	67.98%	73.30%	63.06%	57.58%	58.83%	61.93%	60.63%	56.70%	58.74%	62.50%	66.67%	55.56%
	Not Clear	40.00%	36.63%	11.43%	30.05%	26.24%	35.13%	41.21%	38.50%	36.55%	36.65%	38.66%	37.38%	31.25%	27.78%	33.33%
	Misleading	0.00%	0.99%	0.00%	1.97%	0.46%	0.81%	1.21%	2.67%	1.52%	2.71%	4.64%	3.88%	6.25%	5.56%	11.11%
Timing	< 2 yrs	59.87%	70.27%	66.67%	6.10%	7.84%	11.03%	12.03%	9.13%	9.17%	3.61%	6.88%	6.36%	15.79%	0.00%	7.69%
	2-5 yrs	1.27%	14.41%	19.05%	13.41%	23.14%	28.13%	11.62%	21.53%	26.25%	3.93%	8.70%	6.71%	10.53%	14.29%	7.69%
	> 5 yrs	38.85%	15.32%	14.29%	10.16%	13.33%	13.69%	7.47%	18.61%	15.00%	11.80%	12.32%	12.01%	5.26%	0.00%	15.38%
	Other	–	–	–	70.33%	55.69%	47.15%	68.88%	50.73%	49.58%	80.66%	84.42%	74.91%	68.42%	85.71%	69.23%

Table 3: Distribution across different industry.

4.1 Market Capitalization

Table 2 synthesizes market capitalization data across languages and countries to illustrate the relationship between company size and ESG reporting practices. This analysis confirms existing trends and identifies and discusses contradictory findings that emerge across different linguistic and geographical contexts, providing a more nuanced understanding of this dynamic relationship.

Based on the data, in most languages, including Chinese, French, English, and Korean, high-market-cap companies show a significantly higher rate of explicit promises compared to medium- and low-market-cap firms. For example, in the Chinese dataset, the "Yes" rate for promise identification is 52.78% for high-cap companies, dropping to 23.23% for low-cap firms. This trend suggests that larger corporations are more likely to articulate their ESG commitments explicitly.

This phenomenon is not coincidental but a result of a combination of factors. First, large public companies face increasing pressure from regulators and institutional investors to provide quantifiable, forward-looking ESG commitments to meet compliance and capital allocation needs. Second, high-cap companies have greater financial and human resources to establish dedicated sustainability teams, hire legal experts, and invest in sophisticated reporting infrastructure, enabling them to formulate and track detailed ESG promises. In contrast, smaller firms often lack these resources, which limits their ability to engage in complex ESG reporting. Finally, clear ESG commitments are a vital part of risk management and public relations strategies for large corporations. By actively addressing social and environmental concerns, they aim to mitigate reputational risk and build public trust, which is crucial for globally recognized brands.

In terms of evidence provision and promise clarity, the data present a key point of contradiction. The relationship between market capitalization and clarity is not consistent and varies in language contexts. Although Chinese data show the highest clarity for small firms (80.49% for low-cap companies), Japanese, French, and English data do not follow this trend. Japanese and Korean data show high-cap companies scoring highest in evidence provision and clarity, while French data indicate the highest clarity for medium-cap firms. For the French data, all companies exhibit a high proportion of promises; however, data reveal a significant difference between medium and smaller companies versus larger ones in terms of evidence provision and clarity. This supports the observation in the French and English data that larger firms tend to focus on long-term strategies with less emphasis on tangible achievements than smaller companies.

This discrepancy suggests that the relationship between company size and the clarity of the ESG reporting is not a universal trend but is highly dependent on regional and cultural factors. Overall, the data reveal a trade-off between "clarity and

quantity." Large firms, in a bid to address a wide range of stakeholder concerns, tend to issue a high volume of promises, which can dilute the clarity of individual commitments. Conversely, smaller, less-resourced firms may formulate more targeted commitments that are easier to substantiate and, consequently, clearer. Regarding the timing of commitments, small and medium-sized Taiwan firms show a strong preference for short-term verification (within two years), while large companies have a more balanced distribution, with a significant portion of their promises extending beyond five years. This trend is also observed in other languages.

This temporal difference reflects a core strategic and operational distinction between firms of different sizes. Smaller companies need short-term, verifiable results to quickly demonstrate a return on ESG investments to attract funding and talent. Long-term goals, which require sustained and significant financial outlays, are a more viable option for high-cap corporations that pursue large-scale, capital-intensive ESG projects, such as developing new sustainable technologies. Their strategic planning is inherently long-term, and their ESG commitments reflect this reality.

4.2 Industry-Wise Analysis

Table 3 explores how the dynamics of specific industries influence the formulation, communication and substantiation of the ESG commitments. By consolidating industry data across all languages into a single comprehensive analysis, we can identify overarching sectoral trends and compare reporting nuances within the same industry in different linguistic contexts.

The data reveal that the Energy sector consistently has a high percentage of promises across all analyzed languages (English: 80.12%; French: 78.70%; Japanese: 91.99%; Chinese: 58.24%). The high promise identification-to-actionable evidence ratio of the French and English data suggest a strong reliability of this industry's reports independently from company size. This is further supported by the high proportion of clear evidence, likely reflecting the heightened expectations for innovation and scientific evidence from investors in this field. The same observation is made with the Luxury industry possibly due to its reliance on carefully crafted brand images and the resulting need to redefine practices within the sector. In contrast, the Biomedical industry in the Chinese dataset shows a notably low rate of promise articulation (18.11%).

Despite the low promise rate in the Taiwan Biomedical sector, it boasts an exceptionally high clarity rate of 88.57%. As a matter of fact, the Biomedical sector is a highly regulated field focused on scientific and medical precision, where public statements are subject to intense scrutiny to avoid legal and ethical liabilities, resulting in a lower number of promises that are more precise and well-documented. Meanwhile, by actively articulating a high volume of commitments, companies from the Energy sector aim to demonstrate their commitment to transition, manage reputational risk, and align with global sustainability goals. Overall, this trend indicates that an industry's propensity to issue ESG promises is directly correlated with the level of public and regulatory scrutiny it faces.

In terms of commitment timelines, the Chinese and Japanese Energy sectors show a strong preference for short-term verification. However, the Chinese Semiconductor industry has a significant proportion of long-term commitments (38.85% for more than five years). Conversely, the Finance industry clearly leads with the one of the highest scores for already verifiable actions (Other) exposing the short-term nature of the financial sector initiatives, regardless of company size. These examples epitomize the fact that temporal trends reflect the nature of their business models. In other words, it indicates that the timing of ESG commitments is deeply integrated with an industry's operational cycles and strategic needs.

5 Experiment

5.1 Methods

Following the previous studies discussed in Section 2.2, we explore the RAG approach as a proof of concept and design it for the proposed tasks. Specifically, when given an instance, we first retrieve the six most similar samples in the training set. We leveraged Multilingual E5 Text Embeddings (Wang et al., 2024a) to calculate the cosine similarity between target instance and instances from the training set. Then, we provide the top-six examples for the LLM to perform in-context learning (Dong et al., 2022). We use GPT-4o as the base LLM and run all inferences via API with temperature 0.0. Retrieval computations using E5/E5-V embeddings were performed on four A100-80GB GPUs.

The prompt structure used in the experiment follows this order: task description, annotation proce-

Approach	Task	English	French	Chinese	Japanese	Korean
w/o RAG	Promise Identification (PI)	0.842	**0.816**	0.521	**0.670**	**0.849**
	Actionable Evidence (AE)	0.680	**0.746**	0.163	0.720	**0.792**
	Clarity of Promise–Evidence Pair (CPEP)	0.411	0.443	0.569	0.450	0.897
	Timing for Verification (TV)	0.636	0.523	0.317	0.632	0.406
w/ RAG	Promise Identification	**0.866**	0.798	**0.540**	0.659	0.807
	Actionable Evidence	**0.757**	0.732	**0.503**	**0.850**	0.774
	Clarity of Promise–Evidence Pair	**0.467**	**0.487**	**0.628**	**0.465**	**0.939**
	Timing for Verification	**0.693**	**0.601**	**0.469**	**0.684**	**0.571**

Table 4: Experimental Results (F1 Score). The best performance in each language is denoted in **bold**.

dure, definitions, and context with the target paragraph. Specifically, the prompt is structured, which ensures clarity and consistency in the annotation process, as shown in Appendix C.

5.2 Experimental Results

In the experiment, we randomly select 200 instances from each language as the test set, and the remaining instances are used for training. To prevent data leakage, we ensured a balanced sampling of instances from different ESG reports during the random selection process. While some instances may come from the same document, each focuses on a different aspect, making overlaps non-problematic for our evaluation purposes. We use the F1 score to evaluate the performance of each task. Table 4 shows the performance of each task in each language. The lower clarity for French and Japanese and the timing scores for Korean correlate with lower kappa agreement (about 0.6-0.7). The lower Chinese performance may be due to the reliance on tables and figures in Chinese reports.

Next, we discuss the results of the RAG approach. First, the performance of most tasks improves when RAG is adopted. Specifically, for English and Chinese, all tasks perform better when using RAG. Second, RAG enhances performance in estimating the clarity of the promise–evidence pair and inferring the timing for verification, regardless of the language used. These results suggest the usefulness of RAG in these two novel tasks. Furthermore, the findings demonstrate the value of the proposed annotations. With the proposed dataset, the performance of the fine-grained promise evaluation can be improved. Third, although performance in promise identification and actionable evidence identification tasks may slightly decrease in French, Japanese, and Korean, the decreases are minimal (less than 2% in most cases). These results suggest that the method for retrieving and suggesting samples similar to the instance requires refinement for imbalanced Boolean datasets. In the future, we will focus on improving the RAG approach by extracting balanced samples, particularly for minor labels.

RAG	Task	Chinese		Korean	
		Image–Based	Text–Based	Image–Based	Text–Based
w/o	PI	0.530	0.521	0.837	**0.849**
	AE	0.124	0.163	0.812	0.792
	CPEP	0.510	0.569	0.922	0.897
	TV	0.202	0.317	0.201	0.406
w/	PI	**0.580**	0.540	0.843	0.807
	AE	**0.512**	0.503	**0.845**	0.774
	CPEP	0.618	**0.628**	0.893	**0.939**
	TV	0.297	**0.469**	0.330	**0.571**

Table 5: Image–based experimental results. **Bolded** denotes the best performance in each language. Underlined denotes performance with RAG better than that without RAG.

5.3 SemEval-2025 PromiseEval Task Results

We organized *PromiseEval*, the SemEval-2025 shared task on promise verification, using our dataset (Chen et al., 2025). The system described in this paper was released as the official baseline and was evaluated along with the participant submissions. Our system ranked first in Korean and Japanese and second in English, French, and Chinese. These results indicate that the proposed RAG–based approach is effective and competitive across languages. Compared with our RAG–based baseline, CSCU (Leesombatwathana et al., 2025) pursues lightweight discriminative modeling with data augmentation. They contrast zero-/six-shot GPT-4o, E5-based SVMs, and fine-tuned DistilBERT; paraphrase and synthesis augmentation generally help binary tasks, although clarity remains difficult, while six-shot GPT-4o is the strongest but costly. CYUT (Wu et al., 2025) instead couples RAG with structured prompts and Chain-of-Thought on Llama-3.1-70B to guide multistep reasoning. In PromiseEval, CSCU topped the English track, and CYUT took 1st in French, while our RAG system placed second in both languages.

Input	RAG	Task	ROUGE-L
Text	w/o	Promise Extraction	0.012
		Evidence Extraction	0.007
	w/	Promise Extraction	0.101
		Evidence Extraction	0.139
Image	w/o	Promise Extraction	0.190
		Evidence Extraction	0.230
	w/	Promise Extraction	**0.240**
		Evidence Extraction	**0.317**

Table 6: Results of promise and evidence extraction.

6 Follow-up Experiments

6.1 Image–based Experimental Setup

We noticed a significant difference between the Taiwan/Korea reports and the reports from other countries. As shown in Figure 1, the reports from these two countries utilize a large number of graphs instead of textual descriptions. This observation raises the question of whether we could use multimodal LLMs to read PDF files directly instead of relying on extracted text. To explore this, we expand Korean and Chinese annotations for image–based needs to align them with a PDF page and employ GPT-4o to reassess the tasks using an image as input. For RAG, we leveraged E5-V Universal Embeddings (Jiang et al., 2024) to calculate the cosine similarity between target pages and instances from the training set. We retrieve the two most similar samples for RAG.

Additionally, the task can also be formulated in an extractive manner. Instead of only outputting a yes or no, we can also ask models to extract the promise and evidence from the report. We provide additional annotations in the Chinese dataset and experiment in multimodal settings with and without RAG. We use F1 and ROUGE-L (Lin, 2004) for evaluating classification and extraction. Note that ROUGE-L score is used to evaluate extraction performance. We also use 200 instances for test and the remaining for training.

6.2 Image–based Experimental Results

Table 5 presents the performance. First, using GPT-4o with image input reduces performance in three out of four tasks in the Chinese dataset and in two out of four tasks in the Korean dataset. Second, RAG improves the performance of most tasks when using image input. Third, with RAG, the performance in promise identification and actionable evidence identification tasks improves with Chinese image input, and the performance of actionable evidence identification improves with Korean image input. However, for estimating the clarity of the promise–evidence pair and inferring the timing for verification, using text input with RAG remains superior. In summary, our experimental results suggest that image input should be used for PI and AE tasks, while text input is preferable for CPEP and TV tasks. Additionally, RAG performs well regardless of input type.

Table 6 presents the results. These results indicate that the best performance is achieved in the image–based setting with RAG for both promise and evidence extraction. This emphasizes the importance of exploring multimodal input for ESG report understanding.

7 Conclusion

This paper introduces the concept of Promise Verification, a novel task aimed at evaluating the credibility and fulfillment of promises made by corporations. We propose the first multilingual dataset, ML-Promise, to emphasize the importance of assessing corporate environmental and social promises. Our results demonstrate that RAG improves performance, while also showing the potential of multimodal approaches in promise verification. Our annotations will be released under the CC BY-NC-SA 4.0 license. We hope this work serves as a foundation for the robustness of promise verification systems and contributes to greater accountability in corporate and public disclosures.

8 Outlook

While our current study focuses on clarity verification of promise–evidence pairs at the single-report level, corporate promises are often long-term and span across multiple years. As a natural next step, we plan to extend our framework to compliance checking on an annual basis, verifying whether companies provide sufficient information to enable longitudinal tracking of promise-related actions. This effort will be formalized as part of our 2026 Regulatory Compliance Checking (RegCom) mission[4] . Beyond compliance, we envision integrating the results of ML-Promise and RegCom to establish a systematic methodology for long-term tracking and analysis of corporate promises. By 2030, this will allow us to assess both fulfilled and unfulfilled ESG goals, supporting companies in self-evaluation and enabling regulators to monitor

[4] https://regcom.nlpfin.com/

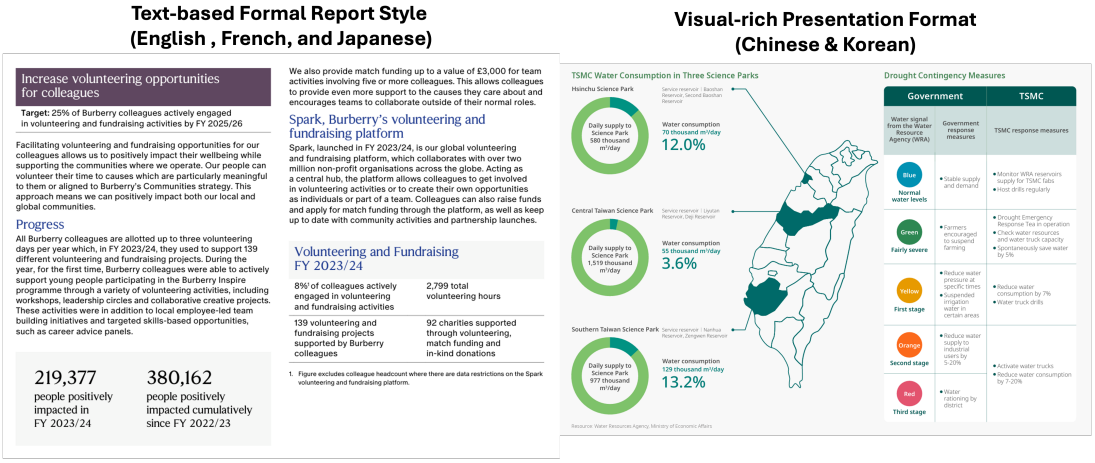

Figure 1: An illustration of our strategy. For readability, all reports here are presented in English.

corporate accountability over extended time horizons.

Limitations

Several limitations warrant discussion. First, although the ML-Promise dataset includes five languages—Chinese, English, French, Japanese, and Korean—its scope is still limited to a few countries and may not fully capture the diversity of corporate promise communication styles globally. The dataset focuses on ESG reports from specific regions, which may limit the generalizability of the findings to other languages and cultural contexts. Future studies can follow our design to expand the dataset to include more regions and languages, which could enhance the robustness and applicability of the proposed methods. In addition, a larger dataset would enhance our results. Second, although the study uses RAG to improve performance, the results show that this approach does not consistently outperform baseline models across all languages and tasks. These inconsistencies suggest that RAG may require further optimization or task-specific adjustments, particularly in handling the nuances of each language and dataset structure. We will also address the need for balanced Boolean labels, particularly in the future improvements for imbalanced datasets. Third, we recognize the dataset's scope is limited to well-documented promises. Note that less publicized or informal commitments are not included in our current scope. Expanding the methodology to incorporate evidence from sources beyond collected documents would enhance coverage. For the real-world challenges, longitudinal verification and diverse contexts should also be taken into account.

These limitations and our findings highlight areas for future research, including expanding the dataset, refining the RAG approach, enhancing multimodal learning, and addressing the inherent ambiguities in corporate ESG reporting.

Ethics Statement

Our study builds ML-Promise exclusively from publicly available ESG reports published by corporations in the U.K., France, Taiwan, Japan, and Korea (reports from 2021 onward), and thus it contains no personally identifiable or sensitive human data; the corpus represents corporate disclosures intended for public communication, and all experiments evaluate models on these materials for research purposes only. We designed multilingual annotation guidelines led by a professional Data & Language Analyst, employed native or work-proficient annotators in each language, and measured substantial inter-annotator agreement; the guidelines, data collection procedures, and competition rules were reviewed by 3DS Outscale legal department, and the annotation study was approved by the University of Tsukuba Institute of Library, Information and Media Science Ethics Review Committee (Notice No. 24-34) as well as by the 3DS Outscale legal department. For the Japanese portion, the fee was (roughly) JPY 150,000 for 150 hours annotation work per annotator. For the French and English datasets, the annotations were carried out by two professional annotators in their capacity as Data and Language Analysts.

Acknowledgment

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

References

Ozgur Ardic, Mahiye Uluyagmur Ozturk, Irem Demirtas, and Secil Arslan. 2024. Information Extraction from Sustainability Reports in Turkish through RAG Approach. In *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

Chung-Chi Chen, Yohei Seki, Hakusen Shu, Anais Lhuissier, Juyeon Kang, Hanwool Lee, Min Yuh Day, and Hiroya Takamura. 2025. SemEval-2025 task 6: Multinational, multilingual, multi-industry promise verification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2461–2471, Vienna, Austria. Association for Computational Linguistics.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anais Lhuissier, Yohei Seki, Hanwool Lee, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024a. 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.

Zhendong Chen, Siu Cheung Hui, Fuzhen Zhuang, Lejian Liao, Meihuizi Jia, Jiaqi Li, and Heyan Huang. 2024b. A syntactic evidence network model for fact verification. *Neural Networks*, 178:106424.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A Survey on In-context Learning. *Preprint*, arXiv:2301.00234.

J. Drchal, H. Ullrich, and T. et al. Mlynář. 2024. Pipeline and dataset generation for automated fact-checking in almost any language. *Neural Comput & Applic*, 36:19023–19054.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. *Preprint*, arXiv:2405.06211.

Dario Garigliotti. 2024. SDG target detection in environmental reports using Retrieval-augmented Generation with LLMs. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 241–250, Bangkok, Thailand. Association for Computational Linguistics.

Nina Gorovaia and Michalis Makrominas. 2024. Identifying greenwashing in corporate-social responsibility reports using natural-language processing. *European Financial Management*.

Marcelo Gutierrez-Bustamante and Leonardo Espinosa-Leal. 2022. Natural Language Processing Methods for Scoring Sustainability Reports? A Study of Nordic Listed Companies. *Sustainability*, 14(15).

Lars Hillebrand, Maren Pielka, David Leonhard, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Milad Morad, Christian Temath, Thiago Bell, Robin Stenzel, and Rafet Sifa. 2023. sustain.AI: a Recommender System to analyze Sustainability Reports. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, pages 412–416, New York, NY, USA. Association for Computing Machinery.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-V: Universal Embeddings with Multimodal Large Language Models. *Preprint*, arXiv:2407.12580.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

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 1935–1947, Vienna, Austria. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual E5 Text Embeddings: A Technical Report. *Preprint*, arXiv:2402.05672.

Shuting Wang, Jiongnan Liu, Shiren Song, Jiehan Cheng, Yuqi Fu, Peidong Guo, Kun Fang, Yutao Zhu, and Zhicheng Dou. 2024b. DomainRAG: A Chinese Benchmark for Evaluating Domain-specific Retrieval-Augmented Generation. *Preprint*, arXiv:2406.05654.

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 494–501, Vienna, Austria. Association for Computational Linguistics.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of Retrieval-Augmented Generation: A Survey. *Preprint*, arXiv:2405.07437.

A Examples

We present illustrative examples and their corresponding annotation results below.[5]

- **Example 1** "Since February 2021, the baskets have been rolled out to BBB stores beginning in *Tsushima* and *Iki*, and are in use at 27 stores as of March 2023. BBB's food-collection boxes made from ocean plastic have been deployed at over 2,000 stores nationwide as of March 2023. We will continue to promote community-based SDGs initiatives."

In this example, the promise is identified as "We will continue to promote community-based SDGs initiatives." The actionable evidence corresponds to the *preceding* segment. The promise–evidence pair is annotated as *Clear*. The verification timeline is labeled as *within 2 years*.

- **Example 2** "CCC will accelerate electrification of powertrains toward carbon neutrality and the introduction of driver-assistance technologies aimed at achieving a society with zero traffic accidents. We are also taking on challenges in new mobility incorporating IoT. To accomplish this goal, CCC will not only pursue the quality of the products and services we offer customers, but also aim to reduce issues at every customer touchpoint."

In this second example, the promise is identified as "CCC will accelerate electrification of powertrains toward carbon neutrality and the introduction of driver-assistance technologies aimed at achieving a society with zero traffic accidents. We are also taking on challenges in new mobility incorporating IoT." The purported actionable evidence appears in the *subsequent* segment; however, because that rationale does not substantiate the stated promise, the promise–evidence pair is annotated as *Misleading*. The verification timeline is labeled as *2–5 years*.

[5]Company names are anonymized as "BBB" and "CCC"; original sentences are lightly summarized.

B Annotation Process

B.1 Annotation Guidelines

The linguistic analysis and the development of common guidelines across multiple languages were led by a professional Data and Language Analyst. This expert collaborated closely with co-organizers to address the unique characteristics of Asian languages and related reports. Each language had native speakers as annotators, ensuring a full understanding of the content during annotation and review.

The annotation guidelines comprehensively outline key aspects of the process, including document type analysis, content evaluation, promise typology classification, and the extraction of promises from visual elements. They provide precise taxonomy definitions (e.g., label descriptions, data segmentation) and core annotation rules to ensure consistency and objectivity. To enhance standardization, the guidelines were developed based on extensive data analysis, identifying recurring patterns to serve as reference points. Annotators followed predefined questions to maintain a consistent approach, such as:

- Should the release date or the evaluation date be used as the time reference?

  From a product commercialization perspective, a potential user scenario involves private auditors assessing a company's transparency and pace of progress based on its initial commitment and target date for achievement. Therefore, the evaluation output should not reflect only the date the report is reviewed by annotators but should be anchored within a clearly defined timeframe — spanning from the company's commitment date to the target date of achievement. Accordingly, the ESG report's release date should serve as the reference point for the evaluation.

- How can consistency be maintained between scientific developments and market ambitions?

  Scientific developments are typically best categorized under the "Longer than 5 years" timeframe (generally spanning 2 to 5 years or more), unless the company indicates a near-term resolution or specifies a target date within 5 years of the report's release. Market ambitions are more appropriately placed within the 2 to 5 year range, as most strategic initiatives aimed at shaping a company's image require a certain amount of time to materialize.

- Should evaluations always consider the longest relevant timeframe?

  Annotators have considered the longest relevant timeframe especially in the three following cases:

  - New Target After Initial Goal Completion: When part or all of the initial goal has been achieved and a new objective with a separate deadline is set, it reflects the company's intent to pursue continuous improvement, annotators assigned the Verification Timeline Tag according to the the final, overarching objective's deadline — whether it's a specific date or a general timeframe.

  - Multiple Measures with Different Timelines: If several actions are taken toward fulfilling the same promise, then the commitment can only be considered fully met when the longest action is completed. Averaging the durations would undermine the integrity of the assessment.

  - Early Measures Within a Broader Transition: When some measures are already verifiable but the original statement clearly indicates an ongoing transition or the beginning of a longer-term project, annotators assigned the tag based on the time required to complete that broader transformation — not on what's already been achieved. This follows the same logic as assessing based on a newly set goal.

- How should the balance between quantity and quality be assessed when evaluating evidence?

  Clarity is determined by both the quantity and quality of evidence. A large amount of evidence does not guarantee clarity — especially if the information is vague, superficial, or only loosely connected to the core promise. In such cases, even a long list of weak or misleading evidence still results in a "Not Clear" or "Misleading" assessment. Quality, in this context, refers to specificity and detail — such as the inclusion of timelines, figures, percentages, identification of involved parties, or clear descriptions of the intended outcome. Ultimately, whether the text is long or short, the key factor is whether the evidence presented is meaningful or merely superficial.

- How should cases be handled where multiple pieces of evidence linked to one promise vary in clarity?

  Since clarity refers to the evaluation of the promise—evidence pair, when multiple pieces of evidence with different clarity levels are present, annotators are instructed to assign the tag corresponding to the *lowest* clarity level. For instance, if even one piece of evidence falls under the "Not Clear" category, the "Not Clear" tag should be applied. The guiding principle is that a single unclear element can reduce the overall clarity, and the model is designed to flag any potential greenwashing for the user to investigate further if desired. As a result, averaging tags or weighting the evidence to determine the most representative clarity level is not allowed.

To further ensure objectivity, paragraph-level segmentation was applied, keeping all relevant evidence within a single topic or sub-topic while minimizing unrelated information. Additionally, definitions were refined using semantic correlations and logical frameworks, ensuring clarity and coherence in the annotation process.

The guidelines, competition rules, data collection methods, and procedures were thoroughly reviewed by three hierarchical levels within 3DS Outscale legal department.

B.2 Data Reference

During the annotation process, PDF documents were used as they are. For text–based experiments in Section 5, the text was extracted from the PDFs, while for image–based experiments in Section 6, the PDF documents were used directly as input. We primarily used PDF parsing tools suited to the language of the document. English and French texts were extracted via an in-house annotation tool which uses in-house PDF Extract API, VintaSoft Imaging .NET SDK, VintaSoft PDF .NET Plug-in and VintaSoft Annotation .NET Plug-in. In cases where the document structure was poorly preserved during extraction, we manually corrected or copied the necessary text to ensure quality.

B.3 Annotators

For the French and English datasets, the annotations were carried out by two professional annotators in their capacity as Data and Language Analysts. The process involved one annotator and one annotation manager (or lead annotator) to ensure accuracy and consistency.

The socio-demographic characteristics of annotators are as follows.

- Gender: two females (English, French, Japanese); two females and two males (Chinese); three males (Korean)
- Age range: 20–30 years old (English, French, Japanese, Chinese, Korean)
- Nationality: European, Japanese, Taiwanese, Korean
- Expertise: one ESG expert (English and French); students specializing in economics (Japanese); master's students in the department of finance (Chinese); finance-related undergraduate degrees and current employment in related companies (Korean)

When annotators assigned different labels, we held a discussion session involving the annotators and the annotation manager to reach a consensus. We typically ensured that the total number of annotators, including the annotation manager, was odd. In cases where no consensus was reached during the discussion, the final label was determined by majority vote.

The annotation study was reviewed and approved by the University of Tsukuba's Institute of Library, Information and Media Science Ethics Review Committee (Notice No. 24-34) and by the 3DS Outscale Legal Department.

C Prompt

Table 7 shows the prompt we used in the experiment.

```
Task description:
You are an expert in extracting ESG-related promises and their corresponding
    evidence from corporate reports that discuss ESG matters. Follow the
    instructions below to ensure careful and consistent annotations.

Annotation procedure:
1. You will be given the content of a paragraph.
2. Determine whether a promise is included and indicate:
   - "Yes " if a promise exists .
   - "No" if no promise exists .
3. If "promise_status" is "Yes", provide the following additional information:
   - "promise_string": Extract the exact wording of the promise from the text,
       without modifying any words.
   - "verification_timeline": Indicate when the promise can reasonably be
       verified. Choose from the following:
     - "within_2_years": Results are expected within 2 years.
     - "between_2_and_5_years": Results are expected within 2 to 5 years.
     - "longer_than_5_years": Results are expected after more than 5 years.
     - "N/A": If no promise is present, or the promise has already been
         implemented.
4. If evidence is included (i.e., if evidence_status is "Yes"), also provide the
    following:
   - The specific part of the evidence (quoted verbatim) (evidence_string)
   - The quality of the relationship between the promise and evidence ("Clear", "
       Not Clear", "Misleading", "N/A") (evidence_quality)

Definitions and criteria for annotation labels:
1. promise_status: A promise consists of a statement related to ESG criteria,
    such as a company's principle, commitment, or strategy.
   - "Yes": A promise exists.
   - "No": No promise exists.
2. verification_timeline: The verification timeline is the assessment of when the
     final results of a given ESG-related action can reasonably be observed and
    the statement verified.
   - "within_2_years": ESG-related measures whose results can be verified within
       2 years.
   - "between_2_and_5_years": ESG-related measures whose results can be verified
       in 2 to 5 years.
   - "longer_than_5_years": ESG-related measures whose results can be verified in
        more than 5 years.
   - "N/A": Use when no promise exists, or when the results can be verified.
3. evidence_status: Evidence refers to information most relevant for showing that
     the core promise is being kept (including, but not limited to, concrete
    examples, company measures, and numbers).
   - "Yes": Evidence supporting the promise exists.
   - "No": No evidence supporting the promise exists.
   - "N/A": Use when no promise exists.
4. evidence_quality: Evidence quality is the assessment of the company's ability
    to support its statement with sufficient clarity and precision.
   - "Clear": Sufficient information is provided; what is said is intelligible
       and logical.
   - "Not Clear": Information is missing to the extent that what is said becomes
       only partially intelligible and/or superficial.
   - "Misleading": The evidence, whether true or not, has no clear connection
       with the claim and appears to divert attention.
   - "N/A": Use when no evidence or no promise exists.
(Due to space constraints, we omit three items in important notes.)
The following are annotation examples of texts similar to the one you will
    analyze. Refer to these examples, consider why they have these annotation
    results, and then output your results.
Examples for references: {context}
Analyze the following text and provide results in the format described above: {
    paragraph}
```

Table 7: Prompt used in our experiments.