ML-Promise: A Multilingual Dataset for Corporate Promise Verification

Anonymous ACL submission

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. This work aims to foster further discourse on the accountability of public commitments across multiple languages and domains.

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

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) points 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.

Our key contributions can be summarized as follows: (1) This study is the first to develop a dataset specifically for identifying ESG-related promises and their supporting evidence using ESG reports. By focusing on the clarity of evidence, our work addresses the challenge of greenwashing and contributes to the broader discussion on corporate accountability. (2) We evaluate the effectiveness and limitations of RAG and multimodal approaches for promise verification across five languages, providing insights into their applicability to ESG analysis.

2 Related Work

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. Chen et al. (2024a) conducted shared-task called ML-ESG series, which aimed to estimate how long the effects of certain events or actions taken by a company will last, impacting its ESG scores. 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.

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 in each language. (number of labels (#) and percentages (%))

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). 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.

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. Some of the example reports are shown in Appendix A. 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 a instance[2] in the ESG reports.

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. protocole 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. Reporting involves the incorporation of tables and pie charts constitute numbered evidence for a textual core 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 Industories 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 to the same standards for clarity and compliance. Therefore, we evaluate different industries under uniform criteria while incorporating multi-

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

ple 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 chosen to reflect each country's economic identity, such as 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 Statistics

Finally, we obtained 3,010 instances, i.e., 600 for each language and 10 additional instances in the Chinese dataset.[3] The Cohen's κ agreement (Cohen, 1960; McHugh, 2012) for these tasks is approximately 0.65-0.96, 0.71-0.88, 0.62-0.80, and 0.60-0.89, respectively. More detailed values are described in Section 4.4. 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.

[3]The Chinese annotators unexpectedly labeled 10 additional instances beyond the intended 600. These instances were included as part of the training dataset for the Chinese experiments.

4 Annotation Process

4.1 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?
- How can consistency be maintained between scientific developments and market ambitions?
- Should evaluations always consider the longest relevant timeframe?
- How should the balance between quantity and quality be assessed when evaluating evidence?
- How should clarity be judged for very short segments (1–2 sentences)? Should they be deemed unclear by default?
- How should cases be handled where multiple pieces of evidence linked to one promise vary in clarity?

To further ensure objectivity, paragraph-level segmentation was applied, keeping all relevant evidence within a single topic or sub-topic while mini-

Task	Language				
	C	E	F	K	J
PI	0.800	0.956	0.731	0.648	0.744
AE	0.730	0.876	0.749	0.810	0.710
CPEP	0.790	0.798	0.702	0.688	0.618
TV	0.830	0.894	0.772	0.602	0.705

Table 2: Kappa coefficient between assessors

mizing 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 our company's legal department.

4.2 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.

4.3 Annotators

For the French and English datasets, annotations were conducted by two professional annotators who are also Data and Language Analysts. The process involved one annotator and one reviewer 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 stundents in the department of finance (Chinese); finance-related undergraduate degrees and current emplyment in related companies (Korean)

Note that the annotation task was approved by our university department Ethics Review Board, and also approved by our company legal department.

```
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. (
       promise_status)

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.

Examples for references: {context}
Instruction for annotation:
Analyze the following text and
    provide results in the format
    described above: {paragraph}
```

Table 3: Prompt used in our experiments.

4.4 Inter-annotator Agreement

The kappa coefficient of classification type attribute is shown in Table 2. For all values, the agreement was above 0.6 (Substantial Agreement (Landis and Koch, 1977)), indicating a workable level of agreement among the annotators' annotations.

5 Experiment

5.1 Methods

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) highlights the advantages of RAG systems, particularly their ability to extract domain-specific information. Fan et al. (2024) discusses training strategies for RAG, including independent, sequential, and joint methods, which can be tailored to optimize retrieval and generation for spe-

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**.

cific domains. For Chinese language applications, Wang et al. (2024b) emphasizes 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. Specifically, when given a 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). In our experiment, we use GPT-4o as the base LLM. We also use four A100-80GB GPUs for our experiment.

The prompt structure used in the experiment follows this order: task description, annotation procedure, definitions, and context with the target paragraph. Specifically, the prompt is structured, which ensures clarity and consistency in the annotation process, as shown in Table 3.

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. 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 timing scores for Korean correlate with lower kappa agreement (about 0.6-0.7). The lower Chinese performance may stem from 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 adopting RAG. 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. Additionally, the findings demonstrate the value of the proposed annotations. With the proposed dataset, the performance of fine-grained promise evaluation can be improved. Third, although the performance in promise identification and actionable evidence identification tasks may slightly decrease in French, Japanese, and Korean, the declines 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 future, we will focus on improving the RAG approach by extracting balanced samples, particularly for minor labels.

6 Follow-up Experiments

6.1 Image-based Experimental Setup

We noticed a significant difference between 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

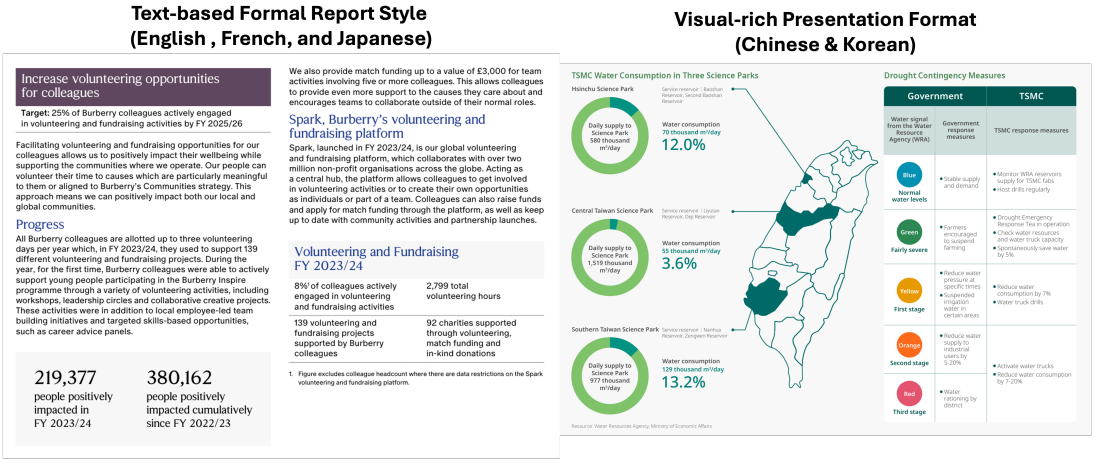

Figure 1: An illustration of our strategy. For readability, all reports here are presented in English.

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.

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.

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 CCBY-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.

Limitation

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 acccount.

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.

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, 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. *arXiv 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, page 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.

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. *arXiv 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.

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 Report Examples

We provide five ESG report examples in this section, and please refer to our training set for more instances: https://drive.google.com/drive/folders/1wWwo5DBY2qFj2KSEqjkjinuK5CB5ku5K?usp=sharing

- English example: https://www.burberryplc.com/content/dam/burberryplc/corporate/2024-updates/burberry-annual-report-and-accounts-2023-24.pdf
- French example: https://www.remy-cointreau.com/app/uploads/2024/04/REMY_COINTREAU_RAPPORT_RSE_2023.pdf
- Chinese example: https://esg.tsmc.com/zh-Hant/file/public/c-all_111.pdf
- Japanese example: https://global.honda/jp/sustainability/report/pdf/2023/Honda-SR-2023-jp-004.pdf
- Korean example: https://kind.krx.co.kr/external/2024/07/24/000126/20240724000069/2023%20POSCO%20Holdings%20Sustainability%20Report%20%28KOR%29.pdf