VeriPromiseESG4K: A Traditional Chinese ESG Promise Verification Dataset for AI CUP 2026

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

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

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

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

Yu-Han Huang
Department of Accounting
National Taipei University
New Taipei City, Taiwan
isyuhann@gmail.com

Ching-Tai Chen
Department of Computer Science
University of Taipei
Taipei, Taiwan
ctchen@utaipei.edu.tw

Yohei Seki
Institute of Library, Information and Media Science
University of Tsukuba
Ibaraki, Japan
yohei@slis.tsukuba.ac.jp

Abstract*—Corporate ESG and sustainability reports have become important information sources for investors, regulators, auditors, and the public. However, these reports often contain broad commitments, future-oriented goals, and policy statements that are difficult to verify without concrete evidence, implementation records, or explicit timelines. Existing ESG NLP datasets mainly focus on relevance detection, topic classification, event analysis, or rating prediction, while resources for evidence-based sustainability promise verification remain limited, especially in Traditional Chinese corporate disclosures. To address this gap, this study introduces VeriPromiseESG4K, a paragraph-level Traditional Chinese dataset for ESG sustainability promise verification. The AI CUP 2026 VeriPromiseESG shared task uses VeriPromiseESG4K v1.0 as its official benchmark dataset. The dataset contains 4,000 annotated paragraphs collected from 2024 ESG or sustainability reports of 50 Taiwanese listed companies. It is designed to support four tasks: promise identification, evidence detection, evidence quality assessment, and verification timeline inference. The central contribution of this work is an iterative human annotation and quality control process for a semantically complex ESG verification task. Twenty-one annotators were organized into seven groups, with each instance independently annotated by multiple annotators. Krippendorff's Alpha was repeatedly calculated to monitor agreement, disagreement cases were reviewed in weekly meetings, annotation guidelines were refined, and a sustainability domain expert performed final adjudication. The final agreement scores demonstrate that the proposed schema can be applied consistently across the four tasks. We also report LLM-based reference baselines with and without retrieval-augmented generation (RAG) to provide an initial benchmark for future AI CUP participants. The baseline results show that the performance gains from RAG are concentrated mainly in verification timeline inference, while promise identification and evidence quality assessment show marginal or mixed changes. Overall, VeriPromiseESG4K provides a new Traditional Chinese ESG NLP resource for sustainability disclosure analysis, promise verification, and greenwashing risk assessment.

***Index Terms*—ESG, sustainability reports, promise verification, Traditional Chinese NLP, dataset construction, human annotation, Krippendorff's Alpha, evidence detection, greenwashing detection**

I. Introduction

Corporate sustainability reporting has become an important channel for firms to communicate ESG performance to investors, regulators, consumers, and the public. As ESG disclosures are increasingly used for investment decisions, regulatory monitoring, supply chain evaluation, and reputation management, their reliability and verifiability have become critical. Companies are expected not only to report ESG goals and policies, but also to provide concrete evidence showing whether these commitments are implemented, measurable, and accountable.

However, sustainability reports often contain broad claims, aspirational goals, and future-oriented commitments without specific actions, measurable indicators, supporting evidence, or clear verification timelines [1]. This makes it difficult for stakeholders to judge whether an ESG promise is substantive, symbolic, or potentially misleading [2]. This issue is closely related to greenwashing, where firms may use environmental or sustainability-related claims to create a positive impression while lacking sufficient evidence of actual performance or implementation [3]. Existing ESG NLP studies mainly focus on topic classification, relevance detection, sentiment analysis, event impact analysis, or rating prediction, but they rarely examine whether corporate sustainability promises are evidence supported and temporally verifiable [4].

To address this gap, this study introduces VeriPromiseESG4K, a paragraph-level Traditional Chinese dataset for ESG sustainability promise verification. The AI CUP 2026 VeriPromiseESG shared task uses VeriPromiseESG4K v1.0 as its official benchmark dataset. The dataset contains 4,000 annotated paragraphs collected from 2024 ESG or sustainability reports of 50 Taiwanese listed companies. It is designed to support four tasks: promise identification, evidence detection, evidence quality

assessment, and verification timeline inference.

In addition to releasing the dataset, this study focuses on how a subjective and semantically complex ESG verification task can be operationalized into a reliable annotation schema. The annotation process was designed as an iterative quality control procedure: annotators independently labeled the same data, the research team calculated Krippendorff's Alpha, disagreement cases were reviewed in weekly meetings, annotation guidelines were revised, and a sustainability domain expert adjudicated disputed instances. This process gradually clarified the decision boundaries among promises, evidence, evidence quality, and verification timelines, producing reliable gold labels for the official shared task.

This work makes three main contributions. First, we release VeriPromiseESG4K, a 4,000-instance Traditional Chinese dataset for ESG promise verification, constructed from the 2024 sustainability reports of 50 Taiwanese listed companies and adopted as the official benchmark of AI CUP 2026. Second, we formalize ESG promise verification as a four-subtask pipeline covering promise identification, evidence detection, evidence quality assessment, and verification timeline inference, and define the corresponding evaluation protocol for the shared task. Third, we present an iterative human annotation and expert adjudication workflow that uses Krippendorff's Alpha to monitor and improve annotation consistency. LLM and RAG results are reported only as reference baselines to indicate task difficulty for future participants.

II. Related Work

ESG-related NLP research has mainly developed along three directions: ESG relevance and topic classification, ESG rating or event impact analysis, and ESG promise verification or greenwashing detection. ESG-FTSE [5] focuses on identifying ESG-related information in news articles, while DynamicESG [6] incorporates temporal news events to analyze ESG rating impact. These resources demonstrate the usefulness of NLP for sustainability analysis, but their primary focus is classification or event impact rather than whether a corporate sustainability promise is supported by specific evidence and can be verified within a concrete time frame.

Greenwashing research has also examined selective disclosure, firm-level behavior, and organizational drivers. Prior studies have proposed typologies of greenwashing measures, deep learning-based greenwashing indexes, and surveys of NLP subtasks for greenwashing detection [1]–[4]. However, many of these approaches either aggregate claims into firm-level indicators or analyze greenwashing from organizational and behavioral perspectives. They do not directly preserve paragraph-level promise–evidence relationships, which are crucial for auditing whether a particular sustainability commitment is concrete, evidence supported, and temporally verifiable.

The most directly related work is ML-Promise [7], a multilingual dataset for corporate promise verification. Although ML-Promise includes a Chinese subset, each language accounts for only a portion of the 3,010 total instances, and the Chinese subset remains relatively limited. VeriPromiseESG4K addresses this gap by providing a dedicated large-scale Traditional Chinese resource of 4,000 paragraph-level instances drawn exclusively from the sustainability reports of 50 Taiwanese listed companies. Unlike news-based ESG datasets, VeriPromiseESG4K is constructed directly from corporate disclosures and emphasizes human-verified annotation quality. The present paper therefore focuses not on proposing a new prediction model, but on dataset construction, annotation schema design, inter-annotator agreement monitoring, and expert adjudication.

III. Dataset

A. Approach Overview

The construction of VeriPromiseESG4K consists of four main stages: data collection, candidate paragraph extraction, human annotation, and agreement-based quality control.

In the first stage, we collected 2024 ESG or sustainability reports from 50 Taiwanese listed companies in the Yuanta Taiwan Top 50 ETF constituents. The companies span major industries including semiconductors, electronics manufacturing, finance, shipping, petrochemicals, steel, telecommunications, retail, biotechnology, and traditional manufacturing.

In the second stage, GPT-5 Pro was used for broad candidate paragraph extraction from the original reports. The extraction process targeted full paragraphs that potentially contained sustainability commitments, goals, future plans, ongoing actions, implementation methods, or supporting evidence. This LLM-assisted step was used only to construct the candidate pool and was not used to determine any official labels.

In the third stage, all official labels, `promise_string`, and `evidence_string` were manually annotated by human annotators using a self-developed web-based annotation platform deployed on Vercel. GPT-5 Pro did not generate final labels or span boundaries.

In the fourth stage, annotation quality was controlled through repeated agreement measurement, disagreement analysis, guideline refinement, and expert adjudication. Krippendorff's Alpha [8] was calculated to evaluate inter-annotator agreement. Disagreement cases were discussed in weekly annotation meetings, and a sustainability domain expert professor made final adjudications. The adjudicated results were used as the gold labels.

B. Data Collection and Official Split

The text in VeriPromiseESG4K mainly comes from 2024 ESG or sustainability reports published by Taiwanese listed companies. We selected 50 companies from the Yuanta Taiwan Top 50 ETF constituents because these companies are highly representative in the Taiwanese capital market, and their sustainability reports are usually more complete and contain richer ESG commitments and implementation information.

Each instance preserves source metadata, including company name, stock ticker, page number, original PDF URL, and company source. The basic unit of the dataset is a full paragraph rather than a single sentence, because commitments

and supporting evidence in corporate sustainability reports are often distributed across different sentences in the same paragraph. Using single sentences as annotation units may break the semantic relationship between a promise and its supporting evidence.

The final dataset contains 4,000 paragraph-level instances, split as shown in Table I.

TABLE I
DATASET STATISTICS AND OFFICIAL SPLIT OF VERIPROMISEESG4K V1.0

Split	Instances	Purpose
Train	1,000	Model training
Validation	1,000	Hyperparameter tuning and model selection
Test	2,000	Final evaluation
Total	4,000	Official benchmark dataset

C. Task Schema and Label Distribution

The four core annotation tasks are `promise_status`, `evidence_status`, `evidence_quality`, and `verification_timeline`. The auxiliary field `esg_type` is provided for analysis but is not included in the official inter-annotator agreement calculation.

Promise identification. `promise_status` determines whether a paragraph expresses a company's commitment to future actions, ongoing actions, policy goals, or improvement directions. Expressions such as "will," "plan," "goal," "commit," or "establish" typically indicate a promise. A promise does not have to be only in the future tense; if an action is currently being implemented as part of a stated commitment, the paragraph can still be labeled as `Yes`.

Supporting evidence detection. `evidence_status` determines whether the promise is supported by concrete evidence. If the paragraph provides only a general vision without concrete actions, systems, timelines, methods, or results, it is labeled as `No`. If the paragraph includes concrete plans, responsible units, system names, quantitative results, or actual records, it is labeled as `Yes`.

Evidence quality assessment. `evidence_quality` evaluates whether the evidence is clear, specific, and verifiable. Labels are `Clear`, `Not Clear`, `Misleading`, and `N/A`. The `N/A` label is used when no promise or no supporting evidence exists.

Verification timeline inference. `verification_timeline` infers when the promise can be verified, using the 2024 report publication year as the starting point. Labels are `already`, `within_2_years`, `between_2_and_5_years`, `more_than_5_years`, and `N/A`. The `N/A` label is used when no promise exists in the paragraph.

Table II summarizes the official task labels and their distributions. Promise identification and supporting evidence detection are both skewed toward the positive class. For tasks where the corresponding promise or evidence is not applicable, `N/A` is included as an official label and its percentage is reported together with the other labels.

TABLE II
OFFICIAL VERIPROMISEESG TASK LABELS AND DISTRIBUTION (%)

Task	Label	%
Promise identification	`Yes`	81.3
	`No`	18.7
Supporting evidence	`Yes`	67.0
	`No`	14.3
	`N/A`	18.7
Evidence quality	`Clear`	55.6
	`Not Clear`	11.3
	`Misleading`	0.1
	`N/A`	33.0
Verification timeline	`already`	36.7
	`within_2_years`	1.9
	`between_2_and_5_years`	24.5
	`more_than_5_years`	18.1
	`N/A`	18.7

Note. `N/A` is included in the reported percentages for applicable tasks.

D. Iterative Annotation and Quality Control

The annotation of VeriPromiseESG4K was completed by 21 annotators from diverse academic backgrounds, including public finance, statistics, accounting, biomedical informatics and medical engineering, computer science, information management, and biomedical-related fields. The 21 annotators were divided into seven groups of three, and each group was assigned data from related industry domains. Each instance was independently labeled by multiple annotators before discussion or adjudication.

The key challenge in constructing the dataset was not only collecting ESG report paragraphs, but also establishing reliable decision boundaries for semantically ambiguous labels. In early annotation rounds, disagreements often arose from three sources. First, annotators sometimes differed on whether a paragraph describing current ESG practice should be considered a promise. Second, evidence quality judgments varied when a paragraph contained concrete-sounding words but lacked measurable outcomes. Third, verification timeline labels were difficult when a paragraph mixed past implementation records, current actions, and long-term future targets.

To address these issues, the research team followed an iterative quality control procedure. After each annotation round, Krippendorff's Alpha was calculated for the four core tasks. Low-agreement cases were collected and reviewed in weekly meetings. During these meetings, annotators and the expert professor discussed ambiguous cases, clarified label definitions, and revised the annotation guidelines. The refined guidelines were then applied to subsequent annotation rounds. This cycle of independent annotation, agreement calculation, disagreement review, guideline refinement, and expert adjudication gradually improved consistency across annotators and produced the final gold labels.

Figure 1 illustrates the human annotation and expert adjudication workflow.

E. Inter-Annotator Agreement

We used Krippendorff's Alpha to measure inter-annotator agreement because it is suitable for multi-annotator and multi-

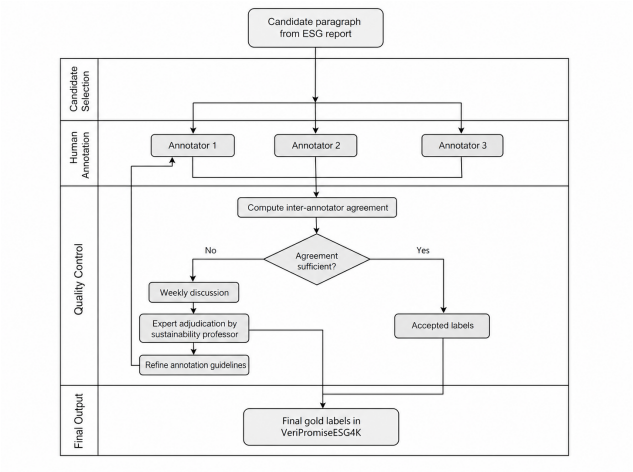

Fig. 1. Human Annotation and Expert Adjudication Workflow for VeriPromiseESG4K

class annotation tasks. Table III reports the final agreement results for the four official VeriPromiseESG tasks.

TABLE III
INTER-ANNOTATOR AGREEMENT FOR THE FOUR VERIPROMISEESG TASKS

Core Task	**Field**	**Alpha**
Promise identification	`promise_status`	0.977
Supporting evidence detection	`evidence_status`	0.962
Evidence quality assessment	`evidence_quality`	0.940
Verification timeline inference	`verification_timeline`	0.955
Overall average	Four core tasks	**0.958**

Promise identification achieved the highest agreement at 0.977, indicating strong annotator consensus on whether a paragraph contained a sustainability commitment. Evidence quality assessment had the lowest Alpha at 0.940, which is expected because judging whether evidence is specific, clear, or potentially misleading is a more subjective semantic task. Nevertheless, the high scores across all four tasks show that the annotation guidelines, weekly disagreement review, and expert adjudication process effectively supported reliable annotation.

F. Dataset Availability and Release Statement

VeriPromiseESG4K v1.0 will be released as the official benchmark dataset for the AI CUP 2026 VeriPromiseESG shared task. To support reproducibility while respecting the intellectual property of the source reports, the dataset will preserve source provenance for each instance, including company name, reporting year, page number, and original PDF URL. The accompanying annotation platform code and evaluation scripts will also be released to facilitate broader research reuse and citation. To preserve the integrity of the shared task evaluation, the public release of full labels will follow the official AI CUP 2026 competition schedule. After the conclusion of the shared task, the dataset and code will be made available through the AIdea competition platform (https://www.aidea-web.tw/aicup_veripromiseesg), the official VeriPromiseESG website (https://veripromiseesg.github.io/), and GitHub (https://github.com/veripromiseesg).

IV. REFERENCE BASELINE

A. Baseline Setting

The baseline evaluation is not intended to propose a competitive system or a new model architecture. Instead, it provides reference results for future AI CUP 2026 VeriPromiseESG participants and helps illustrate the difficulty of the four proposed tasks.

Following the experimental setting of ML-Promise [7], we compare two LLM-based settings: (1) **LLM without RAG**, where the model predicts labels using only task definitions and the input paragraph; and (2) **LLM with RAG**, where retrieved annotated examples are included as in-context demonstrations. RAG is used as a reference baseline because retrieved examples can provide additional task-relevant context during prediction [9]. The 2,000 labeled training and validation instances serve as the retrieval pool, and the 2,000 official test instances are used as the prediction target. For each test paragraph, the paragraph and retrieval pool instances are encoded using Multilingual E5 Text Embeddings [10], and cosine similarity is used to retrieve the six most similar examples as task-specific demonstrations.

The tasks are evaluated using Macro F1, Micro F1, and Weighted F1 [11]. Macro F1 treats all classes equally and is suitable for imbalanced label distributions. Micro F1 aggregates all true positives, false positives, and false negatives before computing F1. Weighted F1 weights each class F1 by class proportion:

$$\text{Weighted F1} = \sum_i \frac{n_i}{N} \times \text{F1}(c_i), \tag{1}$$

where n_i is the number of samples in class c_i and N is the total number of samples.

B. Baseline Results

Table IV reports the consolidated baseline results on the official VeriPromiseESG test set. To keep the paper focused on dataset construction and annotation quality, the table summarizes the three overall F1 metrics and the four-task Macro F1 scores.

The results show that RAG improves the overall Macro F1, Micro F1, and Weighted F1 scores. However, the improvement is not uniform across all tasks. The largest gain appears in `verification_timeline`, where Macro F1 increases from 0.3090 to 0.4227. In contrast, `promise_status` slightly decreases in Macro F1, `evidence_status` remains nearly unchanged, and `evidence_quality` shows only a marginal Macro F1 increase while declining on some other metrics. These results suggest that retrieved examples are particularly helpful for timeline inference, but they are not uniformly beneficial for all ESG promise verification subtasks.

The baseline also highlights two difficult aspects of the dataset. First, timeline inference requires anchoring temporal

TABLE IV
BASELINE PERFORMANCE ON THE OFFICIAL VERIPROMISEESG TEST SET

Method	Overall			Macro F1 by Task			
	Macro F1	Micro F1	Weighted F1	PS	ES	EQ	VT
LLM w/o RAG	0.4976	0.6274	0.6231	0.7298	0.5969	0.3549	0.3090
LLM w/ RAG	0.5234	0.6588	0.6498	0.7152	0.5958	0.3599	0.4227

PS: `promise_status`; ES: `evidence_status`; EQ: `evidence_quality`; VT: `verification_timeline`.

expressions to the report year and distinguishing completed, near-term, medium-term, and long-term commitments. Second, evidence quality assessment requires more than recognizing concrete-sounding words; the model must judge whether the evidence is specific, measurable, and verifiable. These challenges make VeriPromiseESG4K a meaningful benchmark for future research on ESG promise verification.

V. CONCLUSION

This study presented VeriPromiseESG4K, a paragraph-level Traditional Chinese ESG promise verification dataset constructed from the 2024 sustainability reports of 50 Taiwanese listed companies. Unlike existing ESG NLP datasets that mainly focus on topic classification, relevance detection, event impact analysis, or rating prediction, VeriPromiseESG4K examines whether corporate sustainability commitments are present, evidence supported, and temporally verifiable.

The main contribution of this paper is the construction of a reliable benchmark dataset through iterative human annotation and quality control. The annotation process involved 21 annotators, independent labeling, repeated Krippendorff's Alpha calculation, weekly disagreement review, guideline refinement, and expert adjudication. The final agreement scores show that the proposed four-task schema can be applied consistently even though ESG promise verification requires subjective judgments about commitment, evidence clarity, and timeline interpretation.

The LLM and RAG experiments are provided as reference baselines rather than as the main methodological contribution. The results show that RAG gains are concentrated mainly in verification timeline inference, while promise identification and evidence quality assessment show marginal or mixed changes. These findings indicate that current LLM-based methods still face substantial challenges in evidence quality judgment and temporal reasoning.

VeriPromiseESG4K can support automated ESG disclosure analysis for investors, regulators, auditors, and sustainability analysts. It may help stakeholders assess whether corporate ESG commitments include concrete evidence and meaningful timelines. This study is limited by its focus on 2024 reports from 50 Taiwanese listed companies, paragraph-level analysis, and the rarity of the `Misleading` class. Future work can expand the dataset across companies, industries, and years, and incorporate cross-page retrieval, document-level reasoning, multimodal analysis, longitudinal comparison, and fine-tuned ESG models.

ACKNOWLEDGMENT

This research was supported in part by the Ministry of Education (MoE), Taiwan, under the AI CUP 2026 National Collegiate Artificial Intelligence Professional Domain Thematic Competition Project. This work was also supported in part by the Industrial Technology Research Institute (ITRI) and National Taipei University (NTPU), Taiwan, under grants NTPU-114A513E01 and NTPU-113A513E01; the National Science and Technology Council (NSTC), Taiwan, under grant NSTC 114-2425-H-305-003-; and National Taipei University (NTPU) under grant 114-NTPU-ORDA-F-004. The authors are grateful to Prof. Chung-Chi Chen, National Institute of Informatics (NII), Japan, whose ML-Promise framework inspired the task design of this study, for his valuable suggestions on the experimental design.

REFERENCES

[1] Á. Lublóy, J. L. Keresztúri, and E. Berlinger, "Quantifying firm-level greenwashing: A systematic literature review," *Journal of Environmental Management*, vol. 373, p. 123399, 2025, doi: 10.1016/j.jenvman.2024.123399.

[2] X. Wang, X. Gao, and M. Sun, "Construction and analysis of corporate greenwashing index: a deep learning approach," *EPJ Data Science*, vol. 14, no. 1, p. 44, 2025, doi: 10.1140/epjds/s13688-025-00562-w.

[3] M. A. Delmas and V. C. Burbano, "The drivers of greenwashing," *California Management Review*, vol. 54, no. 1, pp. 64–87, 2011.

[4] T. Calamai, O. Balalau, T. L. Guenedal, and F. M. Suchanek, "Detecting Greenwashing: A Natural Language Processing Literature Survey," 2025.

[5] P. Mehra *et al.*, "ESG-FTSE: An ESG Dataset for News Classification," in *Proc. Workshop on Economics and Natural Language Processing*, 2022.

[6] Y. Chen *et al.*, "DynamicESG: A Dataset for Dynamic ESG Impact Analysis," in *Proc. ACL*, 2023.

[7] Y. Seki *et al.*, "ML-Promise: A Multilingual Dataset for Corporate Promise Verification," arXiv preprint arXiv:2411.04473, 2024.

[8] K. Krippendorff, "Computing Krippendorff's alpha-reliability," 2011.

[9] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[10] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," arXiv preprint arXiv:2402.05672, 2024.

[11] J. Opitz, "A closer look at classification evaluation metrics and a critical reflection of common evaluation practice," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 820–836, 2024.