JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, XXX-XXX (2022)
DOI: 10.6688/JISE.20220X_38(X).00XX

LLM-Based ESG Data Annotation: Performance Evaluation and Challenges

BOR-JEN CHEN, AND MIN-YUH DAY*
Graduate Institute of Information Management, National Taipei University
New Taipei City, Taiwan
E-mail: s711336103@gm.ntpu.edu.tw; myday@gm.ntpu.edu.tw

The increasing demand for corporate accountability and data transparency has intensified the need for reliable Environmental, Social, and Governance (ESG) reporting. Yet, the inconsistency in disclosure formats and indicator definitions remains a major obstacle to standardization and automated processing. Despite advancements in generative artificial intelligence, few studies have empirically evaluated the real-world performance of large language models (LLMs) in ESG-related data extraction tasks. This study addresses this gap by investigating how well LLMs can annotate structured ESG metrics from unstructured corporate disclosures.

The research evaluates the annotation performance of three models, Mistral 7B, GPT-4o, and Gemini 2.0 Pro, using ESG reports from four leading Taiwanese semiconductor companies and applying the SASB Semiconductor Standards as the reference framework. Through a comparative analysis of AI-annotated results and human-labeled ground truth across nine quantitative indicators, the study identifies both the capabilities and limitations of LLMs. Key findings reveal that while generative AI performs well when disclosures follow standardized tabular formats, challenges persist in unit conversion, scope differentiation, and vague indicator phrasing.

These insights contribute to the development of more reliable AI-assisted ESG reporting systems by emphasizing the importance of report clarity and prompt design. The results are especially relevant for organizations, policymakers, and AI developers aiming to advance automated sustainability reporting pipelines in data-intensive industries.

Keywords: ESG Reporting, Large Language Models (LLMs), AI Annotation, Sustainability Disclosure, Data Consistency

1. INTRODUCTION

Environmental, Social, and Governance (ESG) reporting has become a vital mechanism for companies to demonstrate sustainability performance and respond to stakeholder expectations [1]. However, ESG disclosures often differ significantly in format, terminology, reporting frameworks, and level of detail, creating challenges for data standardization, cross-company comparison, and automated analysis [2]. Inconsistent ESG reporting not only hinders investors' ability to assess corporate sustainability risks but also complicates regulatory compliance and increases auditing costs, posing practical challenges for both companies and stakeholders.

Driven by rising stakeholder expectations and evolving regulatory mandates, more companies and researchers are exploring AI-assisted solutions to improve ESG reporting workflows. Artificial intelligence has shown potential for automating data extraction,

structuring disclosures, and even drafting initial sustainability narratives. As such, there is a growing need for AI systems that can handle inconsistencies and complexities found in real-world ESG disclosures.

Recent advances in natural language processing (NLP) and generative AI technologies [3, 4], particularly the development of large language models (LLMs) [5] , have created new opportunities for automating ESG data extraction and report analysis. These models are capable of interpreting unstructured text, identifying relevant metrics, and transforming fragmented disclosures into structured formats. Several studies have demonstrated the utility of LLMs in extracting ESG-related information and supporting sustainability assessments [6, 7]. However, notable challenges persist due to ambiguous indicator wording, inconsistent data structures, and variation in how information is presented [8].

While earlier research has proposed conceptual tools and application frameworks, there remains limited empirical evaluation of how generative AI models perform in actual ESG annotation tasks. This study addresses that gap by systematically evaluating and comparing the performance of three LLMs in extracting ESG data based on the SASB Semiconductor Standards.

Specifically, this research assesses the annotation accuracy of Mistral 7B, GPT-4o, and Gemini 2.0 Pro by comparing their outputs against human annotations across ESG reports from four leading Taiwanese semiconductor companies. The study aims to answer three research questions: (1) How accurate are these models in ESG data extraction? (2) What common annotation errors can be identified? (3) How can these findings inform future AI-based ESG report generation?

By focusing on the annotation stage, this study lays the groundwork for improving the data extraction processes that precede report generation. Since accurate and consistent data is essential for any AI-generated ESG reporting system, strengthening model performance at this early step is critical for future automation efforts.

This paper is organized as follows. Chapter 2 reviews related work, Chapter 3 explains the research methodology, Chapter 4 presents the results and discussion, and Chapter 5 concludes the study.

2. RELATED WORK

2.1 ESG Reporting Frameworks and Disclosure Challenges

Environmental, Social, and Governance (ESG) reporting has become a central practice for companies to communicate their sustainability performance and respond to increasing expectations from investors, regulators, and the public [9]. To meet these demands, organizations often refer to international ESG reporting frameworks such as the Global Reporting Initiative (GRI), Sustainability Accounting Standards Board (SASB), and the Task Force on Climate-related Financial Disclosures (TCFD) [10]. Among these frameworks, SASB is particularly important for investors due to its focus on industry-specific and financially material metrics.

However, the ESG reporting landscape remains fragmented. Companies frequently adopt multiple frameworks simultaneously, resulting in inconsistencies in terminology, scope, and data presentation [11]. This lack of harmonization complicates cross-

company comparisons and undermines the usability of ESG disclosures for investment and regulatory purposes. Moreover, ESG reporting is often resource-intensive, requiring significant human and technical capacity to collect, verify, and format relevant data [10].

Challenges also arise in the quality of disclosures. Some reports focus more on qualitative narratives than measurable outcomes, limiting their usefulness for data-driven analysis [11]. The credibility of ESG reports can vary depending on whether third-party assurance is involved [12]. These issues highlight the need for more standardized, efficient, and scalable approaches to ESG data processing.

2.2 Generative AI and Large Language Models (LLMs)

Generative artificial intelligence (AI) refers to technologies that create original content, such as text, images, or audio, based on patterns learned from large dataset [13]. Unlike traditional AI focused on classification or prediction, generative AI emphasizes content generation and adaptability across diverse domains.

The evolution of generative AI has been driven by deep learning advancements, from Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) to the introduction of Transformer architecture [14]. These milestones enabled the development of large language models (LLMs) like BERT [15] and GPT series [16] , significantly advancing natural language understanding and generation. Recent LLMs, such as GPT-4, demonstrate strong multitasking and multimodal reasoning abilities [17].

Prompt engineering has emerged as a critical method for directing LLM outputs, particularly in domain-specific applications like ESG data extraction. Structured prompts, including patterns such as templates and context managers, help control the model's responses and improve consistency in tasks like automated reporting [18]. Effective prompt design enables LLMs to better align with task-specific requirements, reducing hallucinations and enhancing extraction precision in structured tasks like ESG reporting.

Despite their potential, LLMs face challenges including hallucinations [19] , domain adaptation difficulties, and high computational demands [20]. These limitations underline the need for careful evaluation when applying generative AI in complex areas such as sustainability reporting.

Building on these advancements, this study leverages prompt engineering to guide LLMs in extracting ESG indicators based on the SASB Semiconductor Standards, demonstrating how domain-specific alignment can enhance the feasibility of AI-assisted ESG reporting.

2.3 LLM Applications in ESG Reporting

Artificial intelligence (AI), particularly natural language processing (NLP) and large language models (LLMs), has gained increasing attention in the domain of ESG reporting. These technologies have been explored to enhance both the analysis and generation of sustainability disclosures, aiming to improve efficiency, accuracy, and transparency.

Recent studies have demonstrated the potential of transformer-based models, such as GPT-3.5 and GPT-4, in automating ESG tasks like extractive question answering, structured information retrieval, and sustainability risk assessment. For example, CHATREPORT utilized LLMs to assess the alignment of sustainability disclosures with the TCFD framework and provide tailored question-answering capabilities to help

stakeholders navigate lengthy documents [21]. Other studies have focused on multi-label classification of ESG content based on SASB standards, using models like GPT-3.5 in conjunction with resampling techniques to improve the balance of labeled training data [22].

In addition to classification and QA, structuring ESG reports for regulatory frameworks has also been explored. GPT4ESG, for instance, improves ESG con-tent categorization by combining GPT-based models with ESG-BERT and word frequency analysis, enabling clearer ESG risk interpretation [23]. More advanced implementations apply retrieval-augmented generation (RAG) and knowledge graph techniques to transform report content into structured triples, facilitating large-scale analysis [6].

Beyond individual tools, recent frameworks have illustrated how AI can support end-to-end ESG assurance workflows, including data extraction, risk identification, and performance evaluation [24]. These developments highlight the broader role of AI not only in analyzing sustainability disclosures, but also in assisting regulatory compliance and assurance processes.

At the same time, challenges persist. Studies have noted issues with inconsistent indicator labeling, ambiguity in terminology, and hallucination in model outputs, especially when applied to complex real-world reports [8, 25]. Furthermore, effective implementation often requires manual oversight and domain-specific fine-tuning, limiting the immediate scalability of such systems in practice.

Table 1 summarizes representative studies on AI applications in ESG reporting, including the techniques used, their main contributions, and the key challenges identified.

Table 1. Overview of AI Applications in ESG Reporting

AI Technique	Contributions	References
GPT-3.5 Turbo, Extractive QA	Automated question-answering in ESG reports, reducing manual effort.	Automatic Question Answering From Large ESG Reports [26]
GPT-4, Extractive QA, TCFD analysis	Evaluates TCFD conformity and provides traceable Q&A to improve transparency.	CHATREPORT: Democratizing Sustainability Disclosure Analysis [21]
RoBERTa, Distil-RoBERTa	Quantifies ESG texts with improved classification accuracy.	Bridging the Gap in ESG Measurement [7]
WizardLM, RAG, Knowledge Graphs	Extracts structured ESG-related insights, improving data representation.	Glitter or Gold? Deriving Structured Insights from Sustainability Reports [6]
GPT-3.5, ESG-BERT, Multi-label classification	Improves multi-label classification accuracy for ESG reports.	Enhancing ESG Reporting Analysis: Leveraging GPT and Resampling Methods [22]
Custom GPT, ESG-BERT, Word Frequency Analysis	Enhance classification accuracy and stakeholder understanding.	GPT4ESG: Streamlining Environment, Society, and Governance Analysis [23]

While prior research has advanced AI-based ESG information retrieval and classification, few studies have systematically evaluated LLMs' performance on structured ESG

data annotation tasks aligned with reporting standards such as SASB. This study addresses this gap by empirically comparing three LLMs, Mistral 7B, GPT-4o, and Gemini 2.0 Pro in annotating ten key environmental indicators from real-world ESG reports. By analyzing model errors and linking them to disclosure characteristics, this research contributes practical insights into the feasibility and limitations of AI-assisted ESG data processing.

3. METHODOLOGY

This study adopts an evaluation-centered research design to assess how large language models (LLMs) perform in extracting structured ESG metrics from corporate sustainability reports.

3.1 Research Design and Framework

This study follows a multi-step process to evaluate the performance of large language models (LLMs) in ESG data extraction from sustainability reports. The re-search workflow consists of the following stages:

1. **Indicator Selection:** Nine key quantitative ESG indicators were selected from the SASB Semiconductor Standards, covering greenhouse gas emissions, energy management, water management, and waste management.
2. **Data Collection:** ESG reports from four major Taiwanese semiconductor companies (TSMC, MediaTek, UMC, ASE) were collected for the year 2023 from each company's official website.
3. **Manual Annotation:** Two human annotators independently extracted and labeled ESG data based on SASB definitions, including units, reporting format, and page references. Annotation consistency was assessed using Cohen's Kappa, and disagreements were resolved through discussion.
4. **AI Annotation:** Three LLMs, Mistral 7B, GPT-4o, and Gemini 2.0 Pro were provided with structured prompts to perform the same annotation task on the nine selected quantitative indicators.
5. **Comparison and Evaluation:** The results of AI annotations were compared with manual annotations to evaluate accuracy, and common annotation errors were identified and categorized.
6. **Analysis and Discussion:** Based on the results, challenges and implications of using generative AI in ESG reporting were analyzed.

The overall research framework and relationships between these six methodological stages are illustrated in Fig. 1, showing the systematic approach to evaluating LLM performance in ESG data annotation tasks.

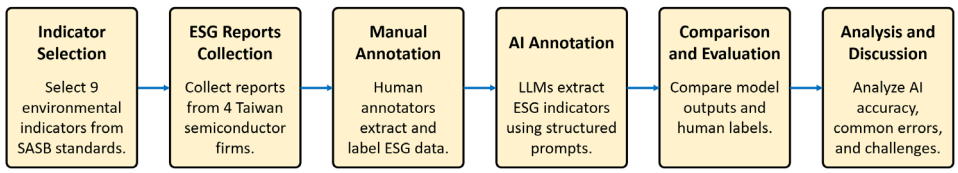


Fig. 1. Research Framework for Evaluating Large Language Model (LLM) Performance in ESG Data Annotation from ESG Reports.

3.2 ESG Report Selection and Indicator Definition

This study focuses on four leading semiconductor companies in Taiwan: Taiwan Semiconductor Manufacturing Company (TSMC), MediaTek Inc., United Microelectronics Corporation (UMC), and Advanced Semiconductor Engineering Inc. (ASE). These companies were selected based on their dominant market positions, global business operations, and relatively mature ESG disclosure practices. As key players in Taiwan's semiconductor industry, which plays a pivotal role in the country's economy and technological development, these firms offer representative cases for evaluating the applicability of AI models in structured ESG data extraction.

To ensure consistency and comparability, this study selected nine quantitative ESG indicators from the Sustainability Accounting Standards Board (SASB) Semiconductor Standards. These indicators were chosen based on their standardized definitions, relevance to environmental impact, and the availability of corresponding data in ESG reports.

The selected indicators span four major environmental themes: greenhouse gas emissions, energy consumption, water usage, and waste management. These serve as the foundation for both manual and AI-based annotation. The indicators are listed in Table 2.

Table 2. ESG Topics and SASB Metrics for the Semiconductor Industry

Topic	Metric	SASB Code
Greenhouse Gas Emissions	(1) Gross global Scope 1 emissions (CO_2-e)	TC-SC-110a.1
	(2) Amount of total emissions from perfluorinated compounds (CO_2-e)	TC-SC-110a.1
Energy Management	(3) Total energy consumed	TC-SC-130a.1
	(4) Percentage grid electricity	TC-SC-130a.1
	(5) Percentage renewable	TC-SC-130a.1
Water Management	(6) Total water withdrawn; percentage in high water stress regions	TC-SC-140a.1
	(7) Total water consumed; percentage in high water stress regions	TC-SC-140a.1
Waste Management	(8) Amount of hazardous waste from manufacturing	TC-SC-150a.1
	(9) Percentage recycled	TC-SC-150a.1

3.3 Annotation Procedures: Manual vs. LLM-Based

This section describes the procedures for both manual and AI-based annotation of ESG disclosures. The process ensures consistency in interpreting and extracting information aligned with the ten selected SASB indicators.

(A) Manual Annotation Workflow

The human annotation process began with a detailed review of nine quantitative SASB indicators across four ESG topics: Greenhouse Gas Emissions, Energy Management in Manufacturing, Water Management, and Waste Management. Each indicator's definition expected unit of measure, and disclosure format were carefully examined according to the official SASB guidelines.

Two human annotators independently labeled ESG disclosures based on these definitions. Each annotation involved recording the reported value, measurement unit, and page reference in a structured spreadsheet.

To assess the reliability of the manual annotations, inter-annotator agreement was calculated using Cohen's Kappa. The resulting Kappa score of 0.857 indicates almost perfect agreement, based on commonly accepted thresholds for interpretation [27].

Most discrepancies were due to variations in interpreting disclosures spread across multiple pages (e.g., differences in rounding, or selecting from multiple formats for the same indicator). These differences were resolved through discussion and consensus prior to conducting the AI evaluation.

(B) Manual Evaluation Criteria

To enhance annotation objectivity, three manual evaluation criteria were defined: data completeness, unit consistency, and presentation format. These criteria guided the manual labeling process and later enabled structured comparison with AI annotations.

Data completeness refers to whether the ESG report discloses the required value(s) for each SASB-defined indicator. An indicator is considered fully disclosed if all required numerical values are provided. For indicators containing two subcomponents, such as those related to water stress regions, partial disclosure is recorded when only one of the two values is reported. If no relevant information is found, the indicator is marked as not disclosed. For example, for the indicator "Total water withdrawn; percentage in high water stress regions," if only total withdrawal is disclosed but the percentage for high water stress regions is missing, the result is marked as partially disclosed.

Unit consistency evaluates whether the reported measurement unit aligns with SASB standards. If the reported unit differs but can be converted (e.g., kWh to GJ), it is marked as requires conversion. If the unit is ambiguous, inconsistent, or missing, it is flagged as unit mismatch. For example, reporting energy in "kWh" instead of "GJ" is "requires conversion.

Presentation format categorizes how the data appears in the report: as a table, paragraph, chart, or hybrid format (e.g., partial paragraph with accompanying image). This classification supports later analysis of how formatting affects AI performance. For example, if "Scope 1 emissions" appears within a structured table with column headers and units, it is labeled as tabular. If it is embedded in text such as "Our Scope 1 emissions reached 1.5 million tons in 2023," it is labeled paragraph.

These evaluation standards ensured consistency between human annotators and provided a clear benchmark for assessing LLM-generated annotations.

(C) AI-based annotation

Three AI models, Mistral 7B, GPT-4o, and Gemini 2.0 Pro, were evaluated for their ability to extract ESG metrics from the companies' reports. Prior to annotation, the relevant chapters and pages that referenced the selected indicators were manually identified and extracted. For Mistral and GPT-4o, the scanned PDF reports were first processed using Adobe OCR and then converted into editable Word documents. These documents were subsequently fed into the models via structured prompts. In contrast, Gemini 2.0 Pro, which supports direct PDF interaction, received the original PDF files along with corresponding prompts. Each model was tasked with extracting the specified ESG metrics based on SASB standard definitions and return numerical values.

(D) Prompt Design for AI Annotation.

To ensure consistency and task relevance across models, a standardized multi-part prompt was developed. This prompt was carefully aligned with the SASB Semiconductor Standards and designed to meet domain-specific annotation requirements. It consisted of the following five components:

1. **Role Definition:** The model was instructed to act as an ESG reporting analyst familiar with the SASB framework, tasked with extracting and annotating sustainability indicators from corporate ESG reports.
2. **Task Overview:** The prompt clarified that the model must extract ten specified environmental metrics based on SASB guidelines, identify the reporting format (e.g., table, paragraph, chart), record page numbers, and note any necessary calculations or missing values.
3. **Indicator List:** A structured table of ten indicators was provided, covering four sustainability topics: greenhouse gas emissions, energy management, water management, and waste management. Each entry included the indica-tor name, type, SASB-defined unit, and official SASB code.
4. **Output Format Requirements:** The model was asked to return outputs in a seven-column table containing: topic, indicator, reporting format, unit, extracted value, page number, and remarks. This format ensured structured comparison across models and against human annotations.
5. **Extraction Rules:** Clear instructions were provided to ensure strict adherence to SASB definitions. These rules emphasized unit conversion where necessary, rejection of hallucinated data, and transparent documentation of any derived values.

The complete prompt was written in Chinese to enhance contextual accuracy and was applied consistently across all three models. The full prompt (translated from the original Chinese) is provided in Appendix A.

4. DATA ANALYSIS AND DISCUSSION

4.1 Manual Annotation Results

This section presents a consolidated analysis of the ESG reports from the four selected companies, focusing on three key dimensions: data completeness, unit consistency, and data presentation format. These factors influence how well AI models can interpret the reports, which is further explored in Section 4.2.

(A) Data Completeness.

UMC demonstrates the highest level of completeness, fully reporting all nine quantitative indicators. TSMC follows with eight fully reported and one partially reported metric, while one is missing. MediaTek and ASE both have eight complete indicators and two partially disclosed ones, with ASE having no missing data.

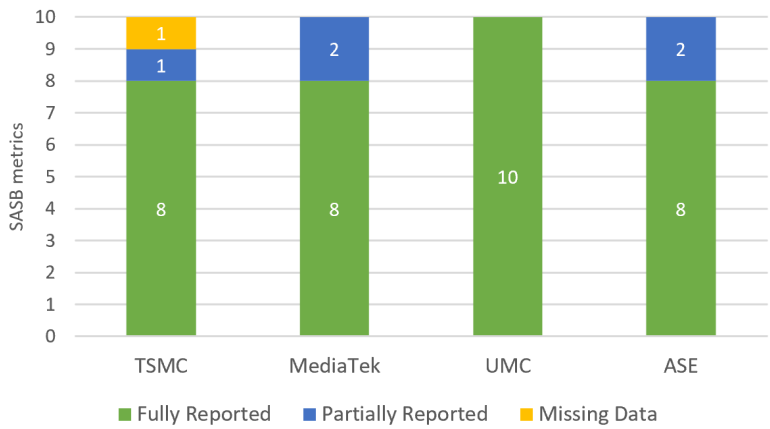


Fig. 2. ESG Reporting Completeness by Company.

(B) Unit Consistency

UMC and ASE maintain complete consistency with SASB-defined units across all indicators. MediaTek requires unit conversion for one metric, while TSMC requires two conversions and contains one case of unit mismatch.

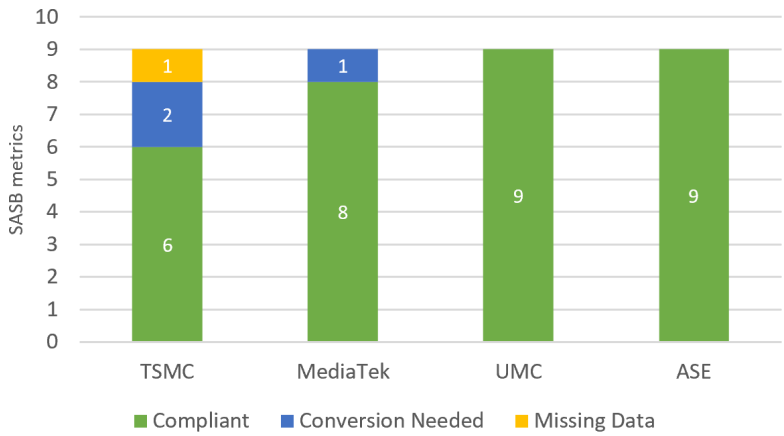


Fig. 3. Unit Consistency in ESG Disclosures.

(C) Presentation Format

UMC and ASE rely predominantly on tabular formats (90%), followed by MediaTek (80%) and TSMC (45%). TSMC shows a more varied structure, including 22% in charts and 33% hybrid formats, potentially impacting AI model accuracy.

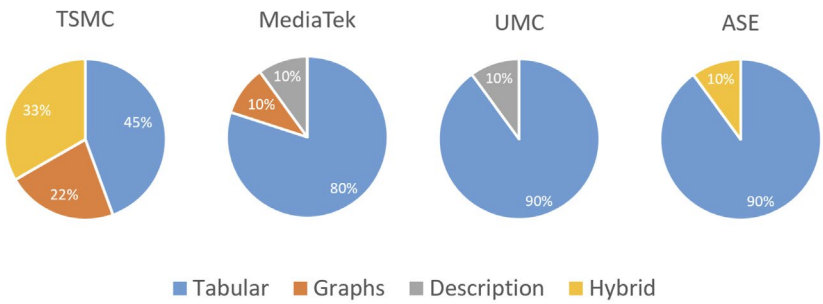


Fig. 4. ESG Data Presentation Format Distribution.

(D) Key Observations

Overall, UMC and ASE demonstrate high reporting quality in terms of data availability and unit standardization, which is likely to support better model performance. TSMC, while reporting most indicators, displays greater diversity in presentation formats, which may introduce ambiguity during AI extraction.

4.2 AI Annotation Performance

This section presents the annotation accuracy of three AI models, Mistral, GPT-4o, and Gemini 2.0, in extracting nine SASB-defined ESG metrics from the reports of four companies. Accuracy is calculated by comparing each model's output to the human-annotated ground truth. An example of the structured annotation output produced by the AI models is shown in Appendix B.

As shown in Table 3, Mistral and Gemini 2.0 achieved the same average accuracy across all companies (80.56%), while GPT-4o demonstrated lower average accuracy (63.89%). Mistral exhibited stable performance across all companies. Gemini 2.0 performed particularly well for UMC and MediaTek, but struggled on ASE's report. GPT-4o showed the highest performance on MediaTek but produced significantly lower accuracy on UMC and ASE.

Table 3. Accuracy of AI Annotations Across Companies.

Company	Mistral	GPT4o	Gemini2.0
TSMC	77.78%	77.78%	77.78%
MediaTek	88.89%	100.00%	88.89%
UMC	88.89%	44.44%	100.00%
ASE	66.67%	33.33%	55.56%

These results indicate that each model may respond differently to structural or contextual variations in ESG reports. While input clarity generally supports better model performance, accuracy is also affected by the model's ability to interpret different formats or indicator labels. The following cases provide further insights into the possible causes behind the variation in model accuracy.

(A) Case 1: Ambiguous Format in TSMC Report

In the TSMC report, the metric *Total Energy Consumed* was disclosed in a mixed format. Part of the data appeared within a narrative paragraph, while related figures were embedded in an adjacent image. Although human annotators successfully identified and combined these pieces of information, all three AI models failed to extract the correct value. This example highlights the challenge AI models face when data is distributed across multiple formats, even when the relevant value is present.

(B) Case 2: Ambiguous Indicator Scope and Labeling in ASE Report

ASE disclosed Scope 1 emissions on multiple pages, with each table referring to different scopes such as site-level or group-level data. However, none of these tables explicitly labeled the value as *total Scope 1 emissions*. Consequently, all three models either selected the incorrect value or marked the indicator as *not disclosed*. This suggests that unclear indicator labeling, and multiple reporting scopes present serious challenges for accurate AI interpretation.

(C) Case 3: Consistent Format but Model-Specific Failure in UMC Report

UMC's ESG report followed standardized naming, consistent SASB units, and clear tabular formats. Both Mistral (88.89%) and Gemini 2.0 (100%) achieved high accuracy for this company. In contrast, GPT-4o only achieved 44.44%, despite the report's clean structure. This case demonstrates that clarity in data presentation alone does not guarantee performance across models. Differences in how models interpret structured prompts or process domain-specific content may lead to unexpected failure. The cause of GPT-4o's underperformance remains unclear and requires further investigation.

(D) Case 4: High Consistency in MediaTek Report

MediaTek's ESG disclosures used consistent tabular formatting and clearly labeled

indicators. All three models performed well on this report, with GPT-4o even reaching 100% accuracy. This case reinforces the notion that a uniform presentation style, aligned with SASB standards, can significantly improve annotation success for different models.

In summary, while report clarity and standardized formatting enhance AI performance, model-specific variability still plays an important role. These case observations suggest that some models, such as GPT-4o, may be more sensitive to hidden factors not captured by formatting alone.

4.3 AI Annotation Statistical Significance Testing

Table 4 presents the results of paired t-tests comparing the annotation accuracy of the three AI models. In all cases, the p-values are greater than 0.05, indicating no statistically significant difference in performance among the models.

This implies that, although some performance variation is observable across companies and metrics, no single model consistently outperforms the others across the dataset. These statistical findings provide a quantitative foundation for the qualitative error analysis presented in Section 4.4.

Table 4. Paired t-test Results Between AI Models.

Comparison	t-statistic	p-value	Interpretation
Mistral vs GPT-4o	1.750	0.088	No significant difference
Gemini 2.0 vs GPT-4o	1.826	0.075	No significant difference
Mistral vs Gemini 2.0	0.000	1.000	No significant difference

4.4 Challenges in AI Annotation

While AI models demonstrate promising potential for ESG data annotation, several recurring error types were observed. These errors help explain the observed differences in model performance across companies and indicators, as summarized in Table 5.

Table 5. Common Annotation Errors by AI Models.

Error Type	Description	Observed Models	Example / Explanation
1. Unit Conversion Errors	AI extracted the correct value but made a mistake in unit conversion or added an unrelated number.	GPT-4o, Gemini 2.0	Converted MWh to GJ incorrectly or appended extra values unrelated to the metric.
2. Wrong Cell in Correct Table	AI located the correct table but selected the wrong cell, often due to complex layouts.	GPT-4o	Picked data from an adjacent row, such as "total water use" instead of "water withdrawn."
3. Scope Confusion	The same indicator was disclosed under different scopes, but AI selected the wrong one.	All models	Extracted emissions for a single facility instead of group-wide total.

4. Indicator Name Ambiguity	Similar wordings led to confusion between different but related indicators.	Mistral, GPT-4o	Mistook "% renewable electricity" for "% renewable energy in total energy consump-
5. Indicator Not Clearly Named	Data was disclosed, but the indicator name was missing or implicit, leading to missed annotations.	GPT-4o	Reported CO_2-equivalent emissions by gas type, but did not explicitly say "total fluorinated emissions."
6. Model Reporting False Negatives	AI model incorrectly marked indicators as "Not disclosed" even when data was clearly present.	GPT-4o	GPT-4o missed 4 disclosed indicators in UMC's report that other models successfully annotated.

Across the three AI models, a range of annotation errors was observed. First, unit conversion errors occurred when models miscalculated units or included irrelevant numerical content. Second, models occasionally selected values from adjacent cells within the correct table, particularly in cases of multi-row or multi-column formatting.

Third, scope confusion occurred frequently, especially when indicators were reported separately by site, business unit, or region. This was notably observed in ASE's report, where the total Scope 1 emissions were not explicitly labeled. Fourth, ambiguous naming introduced uncertainty in cases where multiple indicators used similar terminology, resulting in mismatches between extracted values and actual metrics.

Fifth, even when the underlying data was present, some models failed to detect it due to the absence of clear indicator labeling. These implicit disclosures posed challenges for models without strong contextual inference.

Finally, a unique and notable issue with GPT-4o was the tendency to report false negatives, marking indicators as "Not disclosed" even when they were clearly stated. This was observed in UMC's report, where GPT-4o failed to identify four indicators that were correctly extracted by both Mistral and Gemini 2.0. This behavior suggests that GPT-4o may be more sensitive to formatting nuances or prompt interpretation inconsistencies, particularly in documents with tabular but sparse annotation cues.

In summary, while models generally performed well with clean, tabular data and standard units, ambiguities in wording, inconsistent scope presentation, and lack of explicit labeling reduced overall accuracy. The observed false negatives by GPT-4o also indicate the importance of future work on model robustness and calibration, even when faced with standardized input formats.

5. CONCLUSIONS

This study evaluates the effectiveness of large language models (LLMs) in automating ESG data annotation, using nine quantitative metrics defined by the SASB Semiconductor Standards. By testing Mistral, GPT-4o, and Gemini 2.0 Pro on ESG reports from four leading Taiwanese semiconductor companies, the results demonstrate that while LLMs

generally perform well on clearly structured and consistently formatted disclosures, significant challenges remain.

The evaluation reveals that annotation accuracy is highly dependent on the quality of report formatting. Factors such as unit consistency, indicator clarity, and tabular presentation significantly improve model performance. Conversely, issues such as ambiguous labeling, inconsistent reporting scopes, and implicit indicator descriptions introduce errors. GPT-4o, in particular, exhibited occasional unexplained failures, including false negatives on clearly disclosed indicators, highlighting the need for further model robustness testing.

This study contributes to the growing literature at the intersection of AI and ESG reporting by offering a systematic performance comparison of LLMs in structured data extraction. It identifies six categories of common annotation errors and links them to disclosure characteristics observed during manual annotation. These insights can inform the development of better prompt design strategies, preprocessing tools, and AI-integrated ESG reporting systems.

Furthermore, this research underscores that accurate ESG reporting begins with reliable data extraction. As such, the ability of AI models to correctly interpret and annotate ESG indicators is a foundational step for end-to-end automation. This work lays important groundwork for future systems that not only assist in annotation but also support the generation of comprehensive, auditable ESG disclosures.

Looking ahead, future research may expand this study in several directions. First, the evaluation framework could be applied to other ESG frameworks (e.g., GRI, TCFD) and industries to examine generalizability. Second, experiments with smaller-scale or open-source models may assess performance trade-offs for real-world deployment. Lastly, future work could explore how structured outputs from annotation tasks may feed into narrative generation, enabling a more complete AI-assisted ESG reporting pipeline.

6. ACKNOWLEDGEMENT

This research was supported by the Industrial Technology Research Institute (ITRI) and National Taipei University (NTPU), Taiwan, under grants NTPU-114A513E01 and NTPU-113A513E01; the National Science and Technology Council (NSTC), Taiwan, under grant NSTC 113-2425-H-305-003; the ATEC Group under grant NTPU-112A413E01; and National Taipei University (NTPU) under grant 114-NTPU_ORDA-F-004.

REFERENCES

[1] M. F. Alsayegh, R. Abdul Rahman, and S. Homayoun, "Corporate economic, environmental, and social sustainability performance transformation through ESG disclosure," *Sustainability,* vol. 12, no. 9, p. 3910, 2020.

[2] S. Pizzi, F. Caputo, and E. De Nuccio, "Do sustainability reporting standards affect analysts' forecast accuracy?," *Sustainability Accounting, Management and Policy Journal,* vol. 15, no. 2, pp. 330-354, 2024.

[3] E. Brynjolfsson, D. Li, and L. R. Raymond, "Generative AI at work," National Bureau of Economic Research, 2023.

[4] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative ai," *Business & Information Systems Engineering,* vol. 66, no. 1, pp. 111-126, 2024.
[5] Z. Shen, "Llm with tools: A survey," *arXiv preprint arXiv:2409.18807,* 2024.
[6] M. Bronzini, C. Nicolini, B. Lepri, A. Passerini, and J. Staiano, "Glitter or gold? Deriving structured insights from sustainability reports via large language models," *EPJ Data Science,* vol. 13, no. 1, p. 41, 2024.
[7] T. Schimanski, A. Reding, N. Reding, J. Bingler, M. Kraus, and M. Leippold, "Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication," *Finance Research Letters,* vol. 61, p. 104979, 2024.
[8] L. Mishra *et al.*, "Statements: Universal information extraction from tables with large language models for ESG KPIs," *arXiv preprint arXiv:2406.19102,* 2024.
[9] A. Amel-Zadeh and G. Serafeim, "Why and how investors use ESG information: Evidence from a global survey," *Financial analysts journal,* vol. 74, no. 3, pp. 87-103, 2018.
[10] K. Goswami, M. K. S. Islam, and W. Evers, "A Case Study on the Blended Reporting Phenomenon: A Comparative Analysis of Voluntary Reporting Frameworks and Standards—GRI, IR, SASB, and CDP," *The International Journal of Sustainability Policy and Practice,* vol. 19, no. 2, p. 35, 2023.
[11] S. Arvidsson and J. Dumay, "Corporate ESG reporting quantity, quality and performance: Where to now for environmental policy and practice?," *Business strategy and the environment,* vol. 31, no. 3, pp. 1091-1110, 2022.
[12] P. F. Dilling, P. Harris, and S. Caykoylu, "The Impact of Corporate Characteristics on Climate Governance Disclosure," *Sustainability,* vol. 16, no. 5, p. 1962, 2024.
[13] Y. Cao *et al.*, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226,* 2023.
[14] A. Vaswani *et al.*, "Attention Is All You Need.(Nips), 2017," *arXiv preprint arXiv:1706.03762,* vol. 10, p. S0140525X16001837, 2017.
[15] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, vol. 1: Minneapolis, Minnesota, p. 2.
[16] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165,* 2020.
[17] R. OpenAI, "Gpt-4 technical report. arxiv 2303.08774," *View in Article,* vol. 2, no. 5, 2023.
[18] J. White *et al.*, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382,* 2023.
[19] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems,* 2023.
[20] S. Minaee *et al.*, "Large language models: A survey," *arXiv preprint arXiv:2402.06196,* 2024.
[21] J. Ni *et al.*, "Chatreport: Democratizing sustainability disclosure analysis through llm-based tools," *arXiv preprint arXiv:2307.15770,* 2023.
[22] J.-W. Wu, L. H.-M. Lin, and R. T.-H. Tsai, "Enhancing ESG Reporting Analysis: Leveraging GPT and Resampling Methods for Improved Multi-Label Classification," in *2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, 2024: IEEE, pp. 415-418.

[23] L. H.-M. Lin, F.-K. Ting, T.-J. Chang, J.-W. Wu, and R. T.-H. Tsai, "GPT4ESG: Streamlining Environment, Society, and Governance Analysis with Custom AI Models," in *2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, 2024: IEEE, pp. 442-446.

[24] N. Li, M. Kim, J. Dai, and M. A. Vasarhelyi, "Using artificial intelligence in ESG assurance," *J. Emerg. Technol. Account.,* vol. 21, no. 2, pp. 83-99, 2024.

[25] A. Usmanova and R. Usbeck, "Structuring Sustainability Reports for Environmental Standards with LLMs guided by Ontology," in *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, 2024, pp. 168-177.

[26] P. Parikh and J. Penfield, "Automatic Question Answering From Large ESG Reports," *International Journal of Data Warehousing and Mining (IJDWM),* vol. 20, no. 1, pp. 1-21, 2024.

[27] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics,* pp. 159-174, 1977.

APPENDIX A. INPUT PROMPT TO AI MODELS (TRANSLATED FROM CHINESE)

Role: You are an ESG report analysis expert tasked with annotating ESG disclosures based on SASB standards. Given the provided company report, extract structured environmental indicator data in accordance with **SASB Semiconductor Standards**.

Your Tasks:
- Extract quantitative values aligned with SASB indicators.
- Ensure all extracted values match SASB's standard units.
- Identify the presentation format of each disclosure (Table / Chart / Paragraph).
- Record the page number where the data was found.
- If the value must be calculated, note the calculation method in the Remarks (e.g., subtracting values from two different tables, computing percentages, etc.).
- If data is not disclosed, clearly return "Not disclosed".

SASB Indicators to be Annotated:

SASB Topic	Indicator	Data Type	Unit	SASB Code
Greenhouse Gas Emissions	(1) Total Scope 1 emissions	Quantitative	Metric tons CO_2e	TC-SC-110a.1
Greenhouse Gas Emissions	(2) Total emissions from perfluorinated compounds (PFCs)	Quantitative	Metric tons CO_2e	TC-SC-110a.1
Energy Management in Manufacturing	(1) Total energy consumed	Quantitative	GJ	TC-SC-130a.1
Energy Management in Manufacturing	(2) Percentage grid electricity	Quantitative	%	TC-SC-130a.1
Energy Management in Manufacturing	(3) Percentage renewable energy	Quantitative	%	TC-SC-130a.1
Water Management	(1) Total water withdrawn	Quantitative	m^3	TC-SC-140a.1
Water Management	(2) Total water consumed	Quantitative	m^3	TC-SC-140a.1
Waste Management	(1) Hazardous waste generated	Quantitative	Metric tons	TC-SC-150a.1
Waste Management	(2) Percentage recycled	Quantitative	%	TC-SC-150a.1

Output Format for Annotations: Please output the data in the following table structure:

SASB Topic	Indicator	Format	Unit	Value / Excerpt	Page	Remarks
Greenhouse Gas Emissions	Total Scope 1 emissions	Table	tCO_2e	1,500,000	52	-
Energy Management	Total energy consumed	Paragraph	GJ	3,200,000	78	Converted from 24.7 billion kWh using 1 kWh = 3.6 MJ
Water Management	Total water withdrawn	Table	m^3	50,000,000	115	Summed from Table 3.2 and Table 3.4

Extraction Rules:

1. Strictly follow the SASB indicators. If not disclosed, write “Not disclosed.”
2. If the unit does not match SASB standards, explain the conversion in “Remarks.”
3. If a value is derived through calculation (e.g., summing from different sections), state the method in “Remarks.”
4. Do not infer or generate new data. Only extract explicitly reported content.

APPENDIX B. EXAMPLE AI ANNOTATION OUTPUT (TRANSLATED FROM CHINESE)

SASB Topic	Indicator	Format	Unit	Value / Excerpt	Page	Remarks
Greenhouse Gas Emissions	Total Scope 1 emissions	Table	tCO_2e	1,500,000	52	-
Greenhouse Gas Emissions	Total emissions from PFCs	Table	tCO_2e	Not disclosed	–	PFC emissions not explicitly mentioned in the report
Energy Management	Total energy consumed	Paragraph	GJ	3,200,000	78	Converted from 24.7 billion kWh using 1 kWh = 3.6 MJ
Energy Management	% Grid electricity	Paragraph	%	93.7%	78	-
Energy Management	% Renewable energy	Paragraph	%	60%	78	Target value mentioned in the report
Water Management	Total water withdrawn	Table	m^3	50,000,000	115	Summed from Table 3.2 and Table 3.4
Water Management	Total water consumed	Table	m^3	Not disclosed	–	Not explicitly reported
Waste Management	Hazardous waste generated	Paragraph	t	Not disclosed	–	Not explicitly reported
Waste Management	% Recycled	Paragraph	%	96%	122	Global recycling rate mentioned