

Can Vision-Language Models Assess Environmental Reporting Compliance? A Benchmark for Requirement-Level Disclosure Alignment

Vaishali Aggarwal¹, Ishan Verma¹, Tanay Kumar Gupta¹, Tirthankar Dasgupta²

¹TCS Research, New Delhi, India

²TCS Research, Kolkata, India

{aggarwal.vaishali, ishan.verma, gupta.tanay, dasgupta.tirthankar}@tcs.com

Abstract

Corporate sustainability reports communicate environmental disclosures through text, tables, charts, and infographics, making assessment against sustainability reporting standards challenging. Existing disclosure extraction systems focus on information extraction but do not determine whether disclosed evidence satisfies reporting requirements. We study requirement-level multimodal environmental disclosure alignment, a task that evaluates whether sustainability report evidence satisfies GRI 305 and ESRS E1-4 and E1-6 requirements. To support this task, we curate an expert-annotated benchmark comprising 731 disclosure-alignment instances from 43 sustainability reports. Using this benchmark, we evaluate three Vision Language Models (VLMs): GPT, Gemini, and LLaMA Maverick under three conditioning strategies: open-ended, framework-specific, and knowledge-augmented. Results show that although frontier VLMs possess substantial implicit knowledge of sustainability reporting standards, they frequently produce overgeneralized and optimistic alignment judgments. Requirement-level grounding improves alignment accuracy by 12% and yields more accurate disclosure gap identification. Our findings demonstrate the value of knowledge-grounded multimodal reasoning for scalable, interpretable, and auditable sustainability disclosure assessment.

1 Introduction

Emission disclosures are a central component of corporate sustainability reporting and are increasingly scrutinized by regulators, investors, and auditors (Aiuto et al., 2024). Organizations are expected to report emissions inventories, reduction initiatives, and climate targets in accordance with established sustainability reporting frameworks. Assessing whether reported information aligns with such frameworks, however, remains challenging

because evidence is often distributed across narrative text, tables, charts, and infographics contained within lengthy sustainability reports (Zhang et al., 2025).

Greenhouse Gas Emissions (in mtCO ₂ e)			
	Fiscal Year 2022	Fiscal Year 2023	Fiscal Year 2024
Scope 1 and 2	291,024	278,919	263,330
Scope 1	71,646	65,675	54,917*
Natural Gas	57,867	43,195	36,536
Company Vehicles	7,176	7,742	6,714
Other Fuels	5,952	7,731	5,489
Refrigerants	651	82	119
Process Emissions	not available	6,925	6,059
Scope 2 (market-based)	219,378	213,244	208,413*
Electricity (market-based)	217,643	211,609	206,447
Company Cars	not available	40	159
Purchased or Acquired Heat and Steam	1,735	1,595	1,807

Inputs: GRI Disclosure 305-2 Energy indirect (Scope 2) GHG emissions
Requirements - 1. Gross location-based energy indirect (Scope 2) GHG emissions 2. gross market-based energy indirect (Scope 2) GHG emissions 3. the gases included in the calculation; whether CO₂, CH₄, N₂O, HFCs, PFCs, SF₆, NF₃.

Outputs: (Alignment: Partial, Gaps: 1. Gross location based GHG Emission and 3. Gases included in the calculation data missing)

Figure 1: Example of requirement-level disclosure alignment showing supported and missing reporting requirements from multimodal report evidence.

To improve reporting consistency and comparability, several sustainability frameworks have been developed, including the Global Reporting Initiative (GRI) and the European Sustainability Reporting Standards (ESRS) (Darnall et al., 2022). Although these frameworks address similar environmental topics, they differ substantially in reporting philosophy and granularity (Usmanova et al., 2025). While GRI primarily focuses on indicator level disclosures, ESRS requires detailed methodological explanations, reporting boundaries, target-setting approaches, and assurance-ready evidence. As a result, determining whether a disclosure satisfies framework specific requirements is a non-trivial compliance task.

Recent advances in Large Language Models (LLMs) have enabled increasing research on the automated analysis of sustainability disclosures (Mishra et al., 2024; Sun et al., 2024; Us-

manova and Usbeck, 2024; Goel et al., 2020). Due to the multimodal nature of sustainability reports, recent works have demonstrated the potential of VLMs for interpreting multimodal sustainability reports (Peng et al., 2024). Existing ESG (Environment, Social, and Governance) extraction systems (Anaraki et al., 2025; Wu et al., 2025; Hillebrand et al., 2023; Polignano et al., 2022; Gupta et al., 2024) identify and summarize disclosures, but do not assess whether disclosed evidence satisfies framework requirements at requirement-level granularity. However, compliance assessment requires a fundamentally different capability: determining whether disclosed evidence satisfies explicit reporting requirements. This requires grounding multimodal evidence against framework-specific criteria and identifying missing, incomplete, or weakly supported disclosures. To date, the use of Vision Language Models (VLMs) for requirement-level environmental disclosure alignment remains largely unexplored.

In this paper, we study *requirement-level multimodal environmental disclosure alignment*, a task that evaluates whether evidence contained in sustainability reports satisfies the reporting requirements of GRI 305 (Emissions) and ESRS E1-4 (Targets related to climate change mitigation and adaptation) and ESRS E1-6 (Gross Scopes 1, 2, 3 and Total GHG emissions). To support this task, we curate an expert-annotated benchmark¹ comprising 731 disclosure-alignment instances from 43 sustainability reports spanning 11 industry sectors. Using this benchmark, we evaluate the ability of frontier VLMs under progressively constrained conditioning settings: (i) open-ended setting, which relies on the model’s parametric knowledge, (ii) framework-specific setting, which constrains reasoning to a particular reporting framework, and (iii) requirement-level knowledge-augmented setting, which provides the official disclosure requirements during inference. We further introduce *Coverage Breadth* and *Gap Density* metrics to analyze disclosure completeness and reporting deficiencies at the disclosure level. This work is guided by the following research questions:

- **RQ1:** To what extent can VLMs perform requirement-level environmental disclosure alignment?
- **RQ2:** How do different conditioning settings

influence the accuracy and interpretability of requirement-level alignment judgments?

- **RQ3:** What role does requirement-level knowledge grounding play in reducing over-generalization and improving disclosure-gap identification?
- **RQ4:** To what extent does requirement-level evaluation reveal reporting deficiencies that remain hidden under disclosure-level assessment?

2 Related Work

NLP and VLMs for ESG Analysis: Recent NLP research has increasingly focused on extracting climate and sustainability related information from corporate disclosures. Prior work has explored foundation models for climate-metric extraction (Dave et al., 2024), identification of emission-reduction commitments (ClimateBERT (Schimanski et al., 2023)) and net-zero targets (NetZeroFacts (Wrzalik et al., 2024)), and climate-related QA through retrieval-augmented reasoning (Cao et al., 2025). The growing volume and complexities of sustainability disclosures have motivated LLM-based ESG analysis. Bronzini et al. (Bronzini et al., 2024) combine LLMs with RAG (Lewis et al., 2020) and in-context learning (Dong et al., 2024) to extract structured ESG insights from companies’ sustainability reports for compliance assessment. ChatReport (Ni et al., 2023), ESGReveal (Zou et al., 2025), and recent LLM-based ESG extraction frameworks (Bronzini et al., 2024; Peng et al., 2024; Gutierrez-Bustamante and Espinosa-Leal, 2022) demonstrate strong capabilities for ESG information extraction and structuring. More recently, VLMs have been applied to multimodal sustainability reports (Gupta et al., 2025a; Peng et al., 2024). In particular, Wilhelmi et al. (2026) demonstrate that multimodal reasoning improves the extraction of GHG emission-reduction targets.

ESG Datasets and Benchmarks: With growing importance for ESG and sustainability analysis, several researchers have proposed different benchmark datasets: GRI-QA (Contalbo et al., 2025) for complex table QA, SustainableQA (Ali et al., 2025) for EU taxonomy QA, A3CG (Ong et al., 2025) for aspect action analysis, ClimRetrieve (Schimanski et al., 2024) and ESG-CID (Ahmed et al., 2025) for information retrieval, ESG-Bench (Sun et al., 2026) for hallucination mitigation.

¹<https://github.com/datasharerepo/EnvCompVLM>

Compliance & Framework Alignment: Several studies have explored sustainability framework alignment and compliance-oriented ESG analysis. sustain.AI (Hillebrand et al., 2023) and Anaraki et al. (Anaraki et al., 2025) identify report segments relevant to GRI disclosures, while ESG-KG (Chang et al., 2026) leverages standards-derived knowledge graphs to estimate compliance scores. CLARIESG (Santacroce et al., 2026) combines structured extraction with claim verification for ESG reporting. Gupta et al. (2025b) extract GRI disclosure data and subsequently score the extracted information based on requirements.

Collectively, prior work has advanced ESG information extraction, retrieval, QA, benchmarking, and compliance analysis. However, requirement-level disclosure satisfaction remains largely under-explored, particularly in multimodal settings using VLMs. We address this gap by introducing a benchmark for disclosure alignment and reporting gap identification, and by systematically evaluating frontier VLMs for requirement-level ESG compliance assessment.

3 Reporting Frameworks & Task Formulation

We evaluate disclosure alignment for emissions against two widely used sustainability reporting frameworks: **GRI (305)** and **ESRS (E1-4 & E1-6)**, which differ substantially in reporting philosophy and disclosure granularity.

Standard	Disclosure	Requirements
GRI 305	Direct (Scope 1) GHG Emission 305-1	1. Gross direct (Scope 1) GHG emissions 2. Gases included in the calculation; whether CO ₂ , CH ₄ , N ₂ O, HFCs, PFCs, SF ₆ , NF ₃ 3. Biogenic CO ₂ emissions 4. Source of the emission factors and the global warming potential (GWP) rates used 5. consolidation approach for emissions; whether equity share, financial control, or operational control.
ESRS E1-6	P44	The undertaking shall disclose in metric tonnes of CO ₂ eq its: (a) gross Scope 1 GHG emissions, (b) gross Scope 2 GHG emissions, (c) gross Scope 3 GHG emissions, and (d) total GHG emissions.
	P48	The disclosure on gross Scope 1 GHG emissions required shall include : (a) the gross Scope 1 GHG emissions in metric tonnes of CO ₂ eq; and (b) the percentage of Scope 1 GHG emissions from regulated emission trading schemes.

Table 1: Example reporting requirements from GRI 305 and ESRS E1-6

GRI 305 - GRI 305 adopts an indicator-oriented reporting structure consisting of seven disclosures covering Scope 1, Scope 2, and Scope 3 greenhouse-gas (GHG) emissions, emissions

intensity, emission-reduction initiatives, ozone-depleting substances, and other significant air emissions. Each disclosure is associated with a set of reporting requirements specifying the information that must be disclosed against that particular disclosure.

ESRS E1-4 & E1-6 - In contrast, ESRS E1-4 & E1-6 follow a requirement-oriented structure designed to support assurance-ready sustainability reporting under the Corporate Sustainability Reporting Directive (CSRD). The framework is organized into paragraph-level disclosures, where each paragraph contains one or more reporting requirements. These requirements cover climate targets, methodologies, reporting boundaries, Scope 1-3 emissions and emissions-trading-scheme exposure. Consequently, ESRS places greater emphasis on traceability, methodological transparency, and evidencing disclosure completeness.

Table 1 presents a few examples of how requirements are given in each standard. The complete list of GRI disclosures and ESRS paragraph-level requirements used in our experiments is provided in Appendix B. The differing structures of the two frameworks make the pair a suitable testbed for evaluating whether knowledge-grounded multimodal reasoning can identify missing, partial, and weakly supported environmental disclosures.

Task Formulation: Let f denote a reporting standard framework (GRI or ESRS), D_i^f denote the i -th disclosure under that standard, and $R_i^f = \{r_{i1}, \dots, r_{im}\}$ denote the set of associated reporting requirements. Given a set $I = \{I_1, \dots, I_n\}$ of extracted report pages containing evidence relevant to disclosure D_i^f , the goal is to determine whether the disclosed evidence satisfies the corresponding requirements and to identify any missing information. Formally, the VLM V predicts an alignment label A and a set of disclosure gaps G :

$$A, G = V(I, D_i^f, R_i^f)$$

where

$$A \in \{\text{Full}, \text{Partial}, \text{No}\}$$

and

$$G \subseteq R_i^f.$$

The alignment label is defined as

$$A(I, D_i^f) = \begin{cases} \text{Full} & \forall r_{ij} \in E(I, r_{ij}), \\ \text{Partial} & \exists r_{ij} \in E(I, r_{ij}), \\ \text{No} & \text{otherwise} \end{cases}$$

where $E(I, r_{ij}) \in \{0, 1\}$ denotes evidence identified in the report images corresponding to r_{ij} is sufficient or not.

$$G(I, D_i^f) = \{r_{ij} | r_{ij} \notin E(I, r_{ij})\}$$

The gap set G consists of all requirements for which sufficient supporting evidence is not found in the report images. The gap set provides an interpretable explanation of model decisions by explicitly identifying reporting requirements that lack supporting evidence.

4 Dataset Construction & VLM Conditioning Strategies

This section describes the dataset curation process to support our alignment identification task. We have also shown various VLM conditioning strategies for better alignment and gap identification for interpretable compliance check.

4.1 Benchmark Dataset Construction

Report Collection: We manually collected 56 sustainability reports from SASB repository². After removing incomplete documents, and reports lacking sufficient emissions-related disclosures, the final corpus comprised 43 sustainability reports. The selected reports span all 11 SASB industry sectors, providing diverse reporting practices and levels of disclosure maturity, and environmental-impact profiles. A sector-wise distribution of companies is visualized in Figure 2.

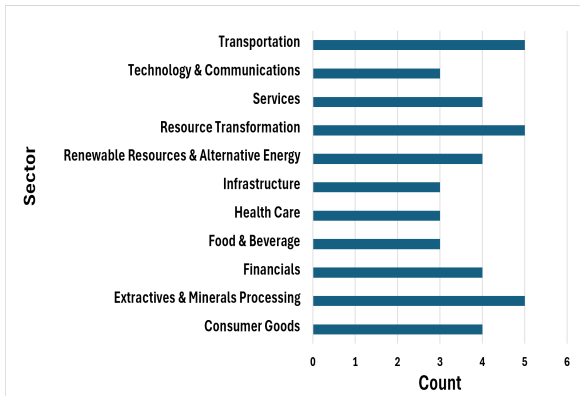


Figure 2: Company count by Sector

Evidence Page Identification: For each report, we first manually searched for GRI or ESRS content-index tables. When disclosure references were explicitly mapped to report pages, the corresponding pages were extracted directly. For reports

²<https://sasb.ifrs.org/company-use/sasb-reporters>

lacking such mappings, we applied a keyword-based retrieval strategy using emission-related terms including *Scope 1*, *Scope 2*, *Scope 3*, *GHG*, *carbon emissions*, and *climate target*. Candidate pages were subsequently reviewed to ensure relevance to the selected disclosure requirements.

Requirement Construction: For each selected disclosure, the corresponding reporting requirements were manually extracted from the official framework documentation^{3,4}. Each GRI disclosure was decomposed into its constituent reporting elements, while ESRS requirements were maintained at paragraph level due to their inherently granular structure. This process produced a set of disclosure-specific requirements that served as the evaluation criteria for alignment assessment.

Annotation Process: To create reference labels, three sustainability compliance experts independently reviewed the extracted report pages and assessed each requirement against the evidence present in the report. At the disclosure level, based on requirements, experts assigned one of the three alignment labels: (i) **Full**: all required information is disclosed. (ii) **Partial**: only a subset of the requirement is satisfied or evidence is incomplete. (iii) **No**: no supporting evidence for the requirement is present.

Alignment Class	Overall	GRI	ESRS
Full	269	121	148
Partial	280	124	156
No	182	56	126
Total	731	301	430
Avg. No. of Pages	3.5 pages per disclosure		

Table 2: Dataset Statistics

Disagreements were resolved through discussion and adjudication. The final annotations constitute the ground-truth alignment labels used for evaluation. The resulting benchmark contains 731 disclosure-alignment instances. Annotation reliability was measured using Fleiss’ Kappa (Fleiss, 1971), yielding a score of 0.81, indicating strong inter-annotator agreement. Each benchmark instance consists of: (i) one or more report pages, (ii) a target disclosure, (iii) the associated reporting requirements and (iv) an expert alignment label.

³<https://www.globalreporting.org/how-to-use-the-gri-standards/gri-standards-english-language/>

⁴<https://xbrl.efrag.org/e-esrs/esrs-set1-2023.html>

4.2 VLM Conditioning Strategies

We evaluate VLMs under progressively constrained evaluation strategies that move from parametric reasoning to requirement-level knowledge grounding:

- **Open-Ended Setting (Parametric Knowledge Only):** Models receive report images and are asked to identify and align emissions-related disclosures without being informed of any specific reporting framework. No external knowledge is provided. This setting evaluates whether VLMs can rely solely on their internal knowledge to recognize and interpret environmental disclosures.
- **Framework-Specific Setting (Constrained Prompting (Yin et al., 2025)):** Models are instructed to perform alignment with respect to a specified reporting framework (GRI or ESRS), but are not provided with the corresponding disclosure requirements. This setting evaluates whether framework awareness alone improves alignment performance while still relying on the model’s parametric knowledge of the framework.
- **Knowledge-Augmented Setting (Requirement Level Injection):** In this setting, models are provided with external knowledge. Instead of relying on models’ internal knowledge, models receive the report images together with the official disclosure requirements associated with the disclosures of the target framework. By grounding predictions in authoritative reporting requirements, this setting enables requirement-level assessment and reduces reliance on implicit framework knowledge. This setup reflects practical compliance-review workflows, where reporting requirements are explicitly available to analysts and auditors.

Together, these settings progressively increase the amount of framework-specific information available to the model, enabling a systematic evaluation of implicit framework knowledge, framework awareness, and requirement-level reasoning. This design allows us to assess the extent to which explicit requirement grounding improves alignment accuracy, interpretability, and disclosure-gap identification.

5 Experiments & Evaluation Metrics

5.1 Models

We evaluate three state-of-the-art Vision Language Models (VLMs): (i) **GPT 5.5** (OpenAI, 2026), (ii) **Gemini 3.5 Flash** (Google), and (iii) **LLaMA 4 Maverick** (Meta AI). These models were selected to represent both proprietary and open-weight multimodal systems with strong document-understanding capabilities. All models are evaluated under identical experimental conditions and receive the same report pages, disclosure definitions, and prompts corresponding to each evaluation setting. Before conducting the alignment experiments, we performed a preliminary capability assessment to verify two prerequisites of the proposed task: (i) comprehension of sustainability-report content, including tables, charts, and narrative text, and (ii) familiarity with reporting-framework concepts and disclosure requirements. The results indicate that all evaluated models demonstrate strong document-understanding capabilities and possess non-trivial knowledge of sustainability reporting standards. Detailed examples and analyses are provided in Appendix C.2 and C.3. Establishing these capabilities allows alignment errors to be attributed primarily to requirement-level reasoning rather than to limitations in visual perception or content extraction.

5.2 Evaluation Metrics

5.2.1 Alignment Based Metrics

The alignment task consists of three ordered labels: $No=0 < Partial=1 < Full=2$. Since these labels exhibit a natural ordinal structure, exact-match evaluation alone does not fully capture the severity of prediction errors. We therefore report complementary metrics that evaluate both classification performance and compliance oriented reasoning behaviour.

Accuracy: Accuracy measures the proportion of instances for which the predicted alignment label exactly matches the expert annotation. While simple and easily interpretable, Accuracy assigns the same penalty to all misclassifications regardless of their distance from the ground-truth label.

Soft Accuracy: To account for the ordinal nature of the labels, we additionally report Soft Accuracy, which assigns partial credit to predictions that differ from the ground-truth label by only one ordinal level. This metric rewards near-correct predictions while still penalizing large deviations. For

each instance, the score is defined as

$$Score_i = \begin{cases} 1.0 & \text{if } A_i^{pred} = A_i^{true}, \\ 0.5 & \text{if } |A_i^{pred} - A_i^{true}| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Soft Accuracy is then computed as

$$Soft_Acc = \frac{1}{N} \sum_{i=1}^N Score_i,$$

where N denotes the number of evaluated instances. Unlike Accuracy, this metric distinguishes near-miss predictions from severe classification errors.

Adjacent Error: The proportion of instances for which the predicted alignment label differs from the ground-truth label by exactly one ordinal level: $|A_i^{pred} - A_i^{true}| = 1$.

Opposite Error: The proportion of instances for which the predicted alignment label differs from the ground-truth label by two ordinal levels: $|A_i^{pred} - A_i^{true}| = 2$.

5.2.2 Analysis Based Metrics

Beyond instance-level alignment evaluation, we introduce two aggregate measures to characterize reporting completeness at the disclosure level: **Coverage Breadth (CB)** and **Gap Density (GD)**. These metrics provide complementary perspectives on disclosure completeness and enable comparisons across reporting frameworks that differ in structure and granularity. Hence, these will be computed on ground truth data.

Coverage Breadth (CB): CB measures the number of disclosures addressed across reports, regardless of whether the alignment is Full or Partial. It quantifies the proportion of reports containing evidence relevant to disclosure D_i^f .

$$CB(D_i^f) = \frac{|\{c : A_c(I, D_i^f) \in \{Full, Partial\}\}|}{N}$$

where N is the number of reports, D_i^f denotes the i^{th} disclosure within framework f , and $A_c(I, D_i^f)$ is the alignment outcome of report c for disclosure i .

Gap Density (GD): While CB captures breadth, GD represents the average proportion of unmet requirements for disclosure D_i^f across reports. It captures the extent to which reporting requirements remain unaddressed even when a disclosure is mentioned:

$$GD(D_i^f) = \frac{1}{N} \sum_{c=1}^N \frac{|G_c(I, D_i^f)|}{|R_i^f|}$$

where $G_c(I, D_i^f)$ denotes the set of missing requirements associated with disclosure D_i^f for report c under framework f and R_i^f is the total number of requirements associated with disclosure D_i^f .

6 Results and Discussions

In this section, we present the overall findings from the three prompting strategies and analyze disclosure-level and sector-level reporting patterns under GRI and ESRS. We further examine the relationship between alignment accuracy, Coverage Breadth (CB), and Gap Density (GD) to understand how reporting completeness influences requirement-level disclosure alignment.

6.1 Findings

We compare three conditioning strategies as mentioned in Section 4, to assess how constraint and grounding change alignment outcomes for GRI and ESRS. Table 3 reports aggregated scores across all disclosures and companies.

Open-Ended (OE): In open-ended setting, all models show limited alignment performance, particularly for ESRS. GPT achieves the strongest overall performance with a combined accuracy of 0.43 and soft accuracy of 0.60, followed closely by Gemini (0.40 and 0.49), while LLaMA performs substantially worse (0.09 and 0.17). The results suggest that frontier VLMs possess some implicit knowledge of sustainability reporting concepts, especially for the widely adopted GRI framework, where all models outperform their ESRS counterparts. This is verified by asking models "give relevant emissions disclosure and their requirements" with framework. Most models provided all disclosures and most of the requirements of the GRI. For ESRS also, they are able to give substantial information. Details are shown in Appendix C. Hence, the outputs skew toward coarse, optimistic judgments and show a bias toward GRI-style disclosures. Additionally, optimism bias remains relatively high, indicating a tendency to overestimate disclosure completeness when no explicit framework guidance is provided.

Framework Specific (FrS) Conditioning: Providing framework-specific instructions (FrS) improves performance across all models. The largest improvement is observed for LLaMA, whose combined accuracy increases from 0.09 to 0.27. GPT remains the strongest model with a combined accuracy of 0.48 and soft accuracy of 0.64. Opti-

		Gemini			Llama			GPT		
		OE	FrS	KA	OE	FrS	KA	OE	FrS	KA
GRI	Accuracy	0.39	0.39	0.5	0.19	0.35	0.56	0.47	0.45	0.61
	Soft Accuracy	0.51	0.53	0.68	0.29	0.47	0.72	0.62	0.63	0.75
	Adjacent Error	0.19	0.25	0.37	0.2	0.24	0.32	0.29	0.35	0.27
	Opposite Error	0.38	0.34	0.13	0.6	0.4	0.11	0.23	0.19	0.11
ESRS	Accuracy	0.41	0.43	0.62	0	0.19	0.69	0.39	0.5	0.67
	Soft Accuracy	0.48	0.57	0.77	0.04	0.38	0.82	0.56	0.65	0.82
	Adjacent Error	0.14	0.25	0.3	0.09	0.23	0.27	0.33	0.29	0.3
	Opposite Error	0.33	0.23	0.09	0.91	0.43	0.04	0.07	0.11	0.03
Combined	Accuracy	0.4	0.41	0.56	0.09	0.27	0.62	0.43	0.48	0.64
	Soft Accuracy	0.49	0.55	0.73	0.17	0.43	0.77	0.6	0.64	0.79
	Adjacent Error	0.17	0.25	0.34	0.14	0.24	0.29	0.31	0.32	0.28
	Opposite Error	0.36	0.29	0.11	0.76	0.42	0.08	0.15	0.15	0.07

Table 3: Performance of evaluated VLMs across GRI, ESRS, and combined benchmark settings under Open-Ended (OE), Framework-Specific (FrS), and Knowledge-Augmented (KA) conditioning.

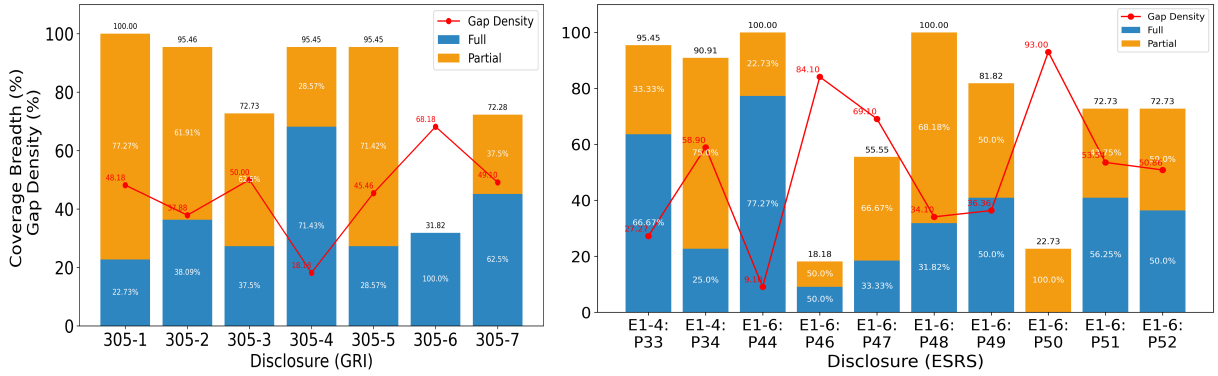


Figure 3: Disclosure wise CB(%) and GD(%) scores

mism bias also decreases across models, showing that framework guidance helps produce more disciplined alignment decisions. However, ESRS performance still lags behind GRI, indicating that framework awareness alone is insufficient for requirement-level assessment.

Knowledge Augmented (KA) Conditioning: Providing the official requirement text and asking for alignment at that granularity flips the pattern and produces the best results across all models and frameworks. GPT obtains the highest combined accuracy (0.64) and soft accuracy (0.79), while LLaMA benefits the most from requirement grounding, improving from 0.09 (OE) to 0.62 (KA). With access to the authoritative requirement text, the model’s decisions became more consistent, requirement-aware, and less susceptible to over-generalisation. Performance gains are particularly pronounced for ESRS, highlighting the importance of explicit requirement-level information for complex reporting standards. Furthermore, optimism bias drops substantially for all models, suggesting that requirement grounding reduces over-generalization and leads to more reliable compli-

ance assessments. A key advantage of knowledge-augmented prompting is that it explains alignment decisions through requirement-level gaps. Instead of producing only an alignment label, the model explicitly identifies unmet reporting requirements, improving interpretability. Figure 1 illustrates how alignment predictions can be directly justified by the corresponding gap set. We observe that, when provided with official requirements, VLMs generate gap sets that are more consistent and closely aligned with expert annotations than those produced under open-ended or framework-specific prompting.

6.2 Disclosure-wise Analysis

Figure 3 presents Coverage Breadth (CB) and Gap Density (GD) across individual disclosures for both frameworks.

GRI- Under GRI 305, most disclosures exhibit high coverage breadth except 305-6, indicating that organizations generally address the required reporting topics. However, the corresponding gap densities reveal notable differences in reporting completeness. For example, 305-1 (Scope 1 emissions)

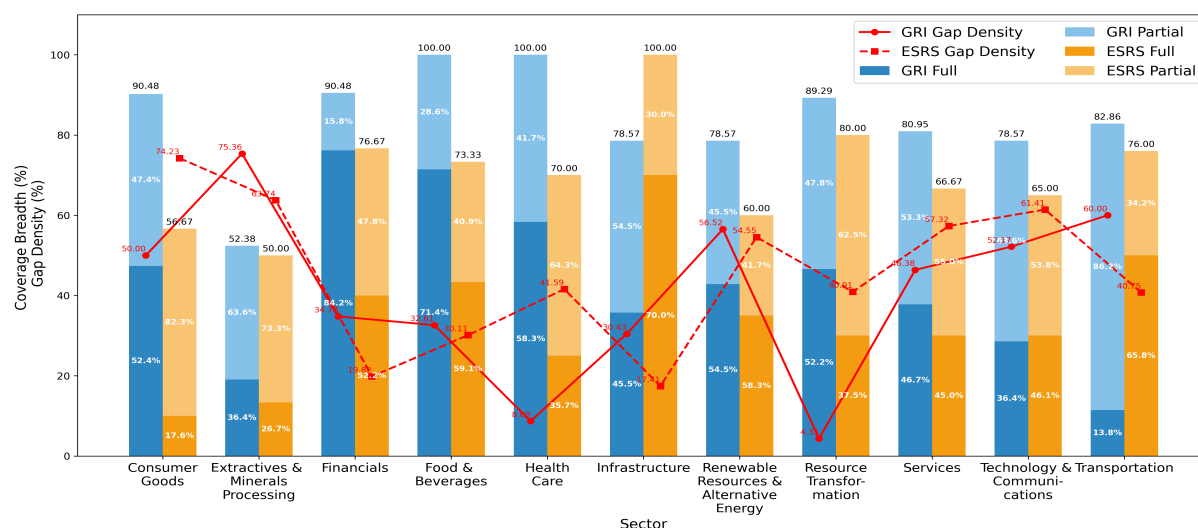


Figure 4: Sector-wise Coverage Breadth (CB%) and Gap Density (GD%) scores

and 305-5 (reduction of GHG emissions) achieve near-complete coverage (100% and 95.45%, respectively), yet maintain moderately high gap densities (48.18 and 45.46), suggesting that disclosures are often only partially complete. For 305-1, the most frequently missing requirements are *biogenic CO₂ emissions* and the *consolidation approach*. In contrast, 305-4 (GHG emissions intensity) combines high coverage breadth with the lowest gap density (18.18), making it the most consistently reported disclosure. Similar patterns are reflected in the alignment metrics (See Appendix C). 305-5 remains one of the most challenging disclosures for models despite its high coverage breadth, reflecting the difficulty of assessing reduction initiatives and supporting methodologies.

ESRS- A similar but more pronounced pattern is observed under ESRS due to its more granular and assurance-oriented requirements. Core quantitative disclosures such as P44 (Scope 1-3 and total GHG emissions) achieve complete coverage breadth (100%) and the lowest gap density (9.1), indicating both broad adoption and high reporting completeness. In contrast, disclosures related to operational control, reporting boundaries, and methodological transparency show the largest reporting deficiencies. P46 exhibits the lowest coverage breadth (18.18%) together with a very high gap density (84.1), while P47 also records high gap density despite moderate coverage. The large proportion of partially aligned disclosures for P48 and P49 further indicates that companies frequently report headline figures but omit supporting methodological details required for full compliance. Consistent

with these observations, models perform best on quantitative disclosures such as P44, P49, P50, and P51 (See Appendix C). In contrast, requirements involving methodologies, assumptions, reporting-boundary changes, and comparability (P34 and P47) consistently obtain lower accuracies across all models.

6.3 Sector-wise Analysis

Sector-wise patterns as shown in Figure 4 reveal substantial differences between disclosure coverage and reporting completeness. Resource Transformation and Financials exhibit high coverage under both GRI and ESRS, but elevated ESRS gap densities indicate missing methodological details, particularly on Scope 2 emissions (E1-6:P49), boundary changes (E1-6:P47), and target disclosures. Extractives & Minerals Processing shows the weakest performance, combining low coverage with high gap densities, reflecting persistent omissions in Scope 3 emissions, pollutant data, and emissions methodologies. HealthCare performs best under GRI, achieving complete coverage with minimal reporting gaps, while Infrastructure demonstrates the strongest ESRS alignment, combining full coverage with the lowest gap density. In contrast, Consumer Goods, Services, Technology & Communications, and Transportation maintain moderate-to-high coverage but relatively high gap densities, driven by recurring omissions in pollutant emissions (GRI 305-7), significant categorywise Scope 3 reporting (ESRS E1-6:P51), target milestones (ESRS E1-6:P47), and decarbonisation-related disclosures (ESRS E1-4:P34). The sector-

wise gap patterns presented in Appendix C.4 (Figures 8a and 8b) further show that these deficiencies are concentrated around a small number of recurring disclosure requirements across sectors.

7 Conclusion

We introduced a requirement-aware formulation for multimodal environmental disclosure alignment and curated an expert-annotated benchmark covering GRI 305 and ESRS E1 requirements. Our experiments show that explicit requirement-level grounding consistently improves alignment performance across VLMs, particularly for the more granular ESRS framework. While the models exhibit some implicit knowledge of sustainability standards, parametric knowledge alone often leads to optimistic and less reliable judgments. In contrast, requirement-level gap identification provides transparent explanations for alignment decisions and supports more interpretable compliance assessment. In future work, we plan to investigate continuous requirement-level alignment scoring instead of discrete labels, develop specialized lightweight models for compliance reasoning, and extend the benchmark to additional sustainability frameworks and environmental topics. We also aim to jointly evaluate alignment quality and gap identification to support comprehensive compliance assessment.

References

- Shafiuddin Rehan Ahmed, Ankit Shah, Quan Hung Tran, Vivek Khetan, Sukryool Kang, Ankit Mehta, Yujia Bao, and Wei Wei. 2025. Enhancing retrieval for esgllm via esg-cid: A disclosure content index finetuning dataset for mapping gri and esrs. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 1–16.
- Kyla Aiuto, Sarah Huckins, and Hannah Momblanco. 2024. What are greenhouse gas accounting and corporate climate disclosures? 6 questions, answered. *World Resources Institute (March 7, 2024)*(www.wri.org/insights/ghg-accounting-corporate-climate-disclosures-explained).
- Mohammed Ali, Abdelrahman Abdallah, and Adam Jatowt. 2025. Sustainableqa: A comprehensive question answering dataset for corporate sustainability and eu taxonomy reporting. *arXiv preprint arXiv:2508.03000*.
- Seyed Alireza Mousavian Anaraki, Danilo Croce, and Roberto Basili. 2025. Unsupervised sustainability report labeling based on the integration of the gri and sdg standards. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 151–162.
- Marco Bronzini, Carlo Nicolini, Bruno Lepri, Andrea Passerini, and Jacopo Staiano. 2024. Glitter or gold? deriving structured insights from sustainability reports via large language models. *EPJ Data Science*, 13(1):41.
- Zhixuan Cao, Ming Han, Jingtao Wang, and Meng Jia. 2025. Carbonchat: large language model-based corporate carbon emission analysis and climate knowledge q&a system. *arXiv preprint arXiv:2501.02031*.
- Li-Yang Chang, Chih-Ming Chen, Hen-Hsen Huang, Ming-Feng Tsai, An-Zi Yen, and Chuan-Ju Wang. 2026. Esg-kg: A multi-modal knowledge graph system for automated compliance assessment. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 602–608.
- Michele Luca Contalbo, Sara Pederzoli, Francesco Del Buono, Venturelli Valeria, Francesco Guerra, and Matteo Paganelli. 2025. Gri-qa: a comprehensive benchmark for table question answering over environmental data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15764–15779.
- Nicole Darnall, Hyunjung Ji, Kazuyuki Iwata, and Toshi H Arimura. 2022. Do esg reporting guidelines and verifications enhance firms’ information disclosure? *Corporate Social Responsibility and Environmental Management*, 29(5):1214–1230.
- Aditya Dave, Mengchen Zhu, Dapeng Hu, and Sachin Tiwari. 2024. Climate ai for corporate decarbonization metrics extraction. *arXiv preprint arXiv:2411.03402*.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Tushar Goel, Palak Jain, Ishan Verma, Lipika Dey, and Shubham Paliwal. 2020. Mining company sustainability reports to aid financial decision-making. In *Proc. of AAAI Workshop on Know. Disc. from Unstructured Data in Fin. Services*.
- Google. Gemini 3.5 flash documentation. <https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash>. Accessed: 2026-07-16.

- Tanay Gupta, Tushar Goel, and Ishan Verma. 2025a. Exploring multimodal language models for sustainability disclosure extraction: A comparative study. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 141–149, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tanay Gupta, Tushar Goel, and Ishan Verma. 2025b. [Leveraging large language models for sustainability disclosures extraction and verification](#). In *Australasian Conference on Information Systems (ACIS 2025)*. Association for Information Systems.
- Tanay Kumar Gupta, Tushar Goel, Ishan Verma, Lipika Dey, and Sachit Bhardwaj. 2024. Knowledge graph aided llm based esg question-answering from news. In *KG4S@ ESWC*, pages 65–78.
- 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):9165.
- Lars Hillebrand, Maren Pielka, David Leonhard, Tobias Deußner, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Milad Morad, Christian Temath, Thiago Bell, and 1 others. 2023. sustain. ai: a recommender system to analyze sustainability reports. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 412–416.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Meta AI. Introducing llama 4: Multimodal intelligence. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>. Accessed: 2026-07-16.
- Lokesh Mishra, Cesar Berrospi, Kasper Dinkla, Diego Antognini, Francesco Fusco, Benedikt Bothur, Maksym Lysak, Nikolaos Livathinos, Ahmed Nasar, Panagiotis Vagenas, and 1 others. 2024. Esg accountability made easy: Docqa at your service. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23814–23816.
- Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stambach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, and 1 others. 2023. Chatreport: Democratizing sustainability disclosure analysis through llm-based tools. *arXiv preprint arXiv:2307.15770*.
- Keane Ong, Rui Mao, Deeksha Varshney, Erik Cambria, and Gianmarco Mengaldo. 2025. Towards robust esg analysis against greenwashing risks: Aspect-action analysis with cross-category generalization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14854–14879.
- OpenAI. 2026. Gpt-5.5 models documentation. <https://developers.openai.com/api/docs/models/gpt-5.5>. Accessed: 2026-07-16.
- Jiahui Peng, Jing Gao, Xin Tong, Jing Guo, Hang Yang, Jianchuan Qi, Ruiqiao Li, Nan Li, and Ming Xu. 2024. Advanced unstructured data processing for esg reports: A methodology for structured transformation and enhanced analysis. *arXiv preprint arXiv:2401.02992*.
- Marco Polignano, Nicola Bellantuono, Francesco Paolo Lagrasta, Sergio Caputo, Pierpaolo Pontrandolfo, and Giovanni Semeraro. 2022. An nlp approach for the analysis of global reporting initiative indexes from corporate sustainability reports. In *Proceedings of the first computing social responsibility workshop within the 13th language resources and evaluation conference*, pages 1–8.
- Marta Santacroce, Michele Luca Contalbo, Sara Pederzoli, Riccardo Benassi, Venturelli Valeria, Matteo Paganelli, and Francesco Guerra. 2026. Clariesg: An end-to-end system for esg analysis over complex tables in corporate reports. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 86–100.
- Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. Climatebert-netzero: Detecting and assessing net zero and reduction targets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15745–15756.
- Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. Climretrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524.
- Siqi Sun, Ben Peng Wu, Mali Jin, Peizhen Bai, Hanpei Zhang, and Xingyi Song. 2026. Esg-bench: Benchmarking long-context esg reports for hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39322–39330.
- Zounachuan Sun, Ranjan Satapathy, Daixue Guo, Bo Li, Xinyuan Liu, Yangchen Zhang, Cheng-Ann Tan, Ricardo Shirota Filho, and Rick Siow Mong Goh. 2024. Information extraction: Unstructured to structured for esg reports. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 487–495. IEEE.
- Aida Usmanova, Rana Abdullah, Debayan Banerjee, Markus Leippold, and Ricardo Usbeck. 2025. Reportgri: Automating gri alignment and report assessment. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6723–6727.

Aida Usmanova and Ricardo Usbeck. 2024. 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)*, pages 168–177.

Lars Wilhelmi, Christian Bruns, and Matthias Schumann. 2026. Enhancing the extraction of ghg emission-reduction targets from sustainability reports using vision language models. *Machine Learning and Knowledge Extraction*, 8(2):37.

Marco Wrzalik, Florian Faust, Simon Sieber, and Adrian Ulges. 2024. Netzerofacts: Two-stage emission information extraction from company reports. 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 70–84.

Qilong Wu, Xiaoneng Xiang, Huang Hejia, Xuan Wang, Yeo Wei Jie, Ranjan Satapathy, Ricardo Shiota Filho, and Bharadwaj Veeravalli. 2025. Susgen-gpt: A data-centric llm for financial nlp and sustainability report generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1184–1203.

Xiaojie Yin, Qilong Wang, and Qinghua Hu. 2025. Constrained prompt enhancement for improving zero-shot generalization of vision-language models. *arXiv preprint arXiv:2508.17417*.

Lei Zhang, Xin Zhou, Chaoyue He, Di Wang, Yi Wu, Hong Xu, Wei Liu, and Chunyan Miao. 2025. Mmes-gbench: Pioneering multimodal understanding and complex reasoning benchmark for esg tasks. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12829–12836.

Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, ZongXiong Lei, Zihan Zeng, Shiming Yang, Hongxiang Tong, Lei Xiao, and Wenwen Zhou. 2025. Esgre-veal: An llm-based approach for extracting structured data from esg reports. *Journal of Cleaner Production*, 489:144572.

A Limitations

This study focuses on environmental disclosures related to GRI 305 and ESRS E1-4,6 requirements. While these frameworks cover important greenhouse-gas reporting obligations, they represent only a subset of sustainability and climate reporting standards. Consequently, the benchmark does not capture the full diversity of environmental, social, and governance disclosures encountered in practice.

The benchmark contains 731 disclosure-alignment instances collected from 43 sustainability reports spanning multiple sectors. The dataset was primarily designed as an evaluation benchmark for requirement-level disclosure alignment and compliance assessment rather than as a large-scale training resource. As a result, its size may limit the direct training of specialized models and is better suited for evaluating reasoning and compliance capabilities of existing VLMs.

Our evaluation focuses on conditioning frontier VLMs under different levels of requirement grounding rather than training task-specific models. Consequently, the findings primarily characterize the reasoning behaviour of existing VLMs when performing requirement-level compliance assessment.

Finally, although requirement-level gap identification enhances the interpretability of alignment decisions, this work primarily focuses on evaluating alignment performance and providing a qualitative analysis of disclosure gaps. Developing standardized quantitative metrics for assessing requirement-gap prediction, as well as extending the benchmark to additional sustainability reporting frameworks, remains an important avenue for future research. Furthermore, the stability and faithfulness of generated gap sets and evidence-based classification decisions warrant systematic investigation to establish the reliability and robustness of the proposed approach.

B Reporting Standards for Environmental Disclosures

GRI Framework: The GRI Standards constitute a modular and interconnected framework designed to enable organizations to publicly disclose the impacts of their activities in a structured and transparent manner. The framework comprises three series of standards, facilitating consistent and comparable sustainability reporting for stakeholders and other interested parties. For environmental disclosure, this study adopts **GRI 305: Emissions** as the relevant standard. GRI 305 encompasses a set of specific disclosures, each addressing a distinct dimension of organizational emissions.

- **GRI 305-1:** Direct (Scope 1) greenhouse gas (GHG) emissions
- **GRI 305-2:** Energy indirect (Scope 2) GHG emissions

- **GRI 305-3:** Other indirect (Scope 3) GHG emissions
- **GRI 305-4:** GHG emissions intensity
- **GRI 305-5:** Reduction of GHG emissions
- **GRI 305-6:** Emissions of ozone-depleting substances (ODS)
- **GRI 305-7:** Nitrogen oxides (NO_x), sulfur oxides (SO_x), and other significant air emissions

ESRS Framework: The European Sustainability Reporting Standards (ESRS), mandated under the Corporate Sustainability Reporting Directive (CSRD), define disclosure requirements across Environmental, Social, and Governance topics. Emission is addressed under **ESRS E1: Climate Change**. Unlike GRI's indicator-based approach, ESRS uses paragraph-level requirements for detailed reporting. For this study, we considered the following Paragraphs:

- **ESRS E1-4: P33:** Disclosure of whether/how GHG reduction targets are set to manage climate risks, impacts, and opportunities.
- **ESRS E1-4: P34:** Comprehensive GHG reduction targets: absolute and intensity values; Scope 1-3 coverage; defined base year and update cycle; 2030/2050 milestones; science-based alignment with methodologies and assumptions; and expected decarbonisation levers with their quantitative contributions.
- **ESRS E1-6: P44:** Gross Scopes 1, 2, 3 and Total GHG emissions-GHG emissions per scope
- **ESRS E1-6: P46:** Inclusion of emissions from associates/JVs/unconsolidated entities based on operational control.
- **ESRS E1-6: P47:** Disclosure of significant changes in definition of what constitutes reporting undertaking and its value chain and explanation of their effect on year-to-year comparability of reported GHG emissions.
- **ESRS E1-6: P48:** Scope 1 emissions and percentage under regulated emissions trading schemes.

- **ESRS E1-6: P49:** Scope 2 emissions reported using both location-based and market-based methods.
- **ESRS E1-6: P50:** Disaggregation of Scope 1 and 2 emissions into consolidated group vs. operational-control investees.
- **ESRS E1-6: P51:** Scope 3 emissions disclosed by each significant (material) Scope 3 category.
- **ESRS E1-6: P52:** Total emissions reported with distinction between location-based and market-based Scope 2 totals.

While GRI disclosures are often organized as indicator-style requirements, ESRS environmental requirements are more granular and assurance-oriented, which increases the risk that high-level reporting narratives appear aligned while missing required breakdowns or evidence. This contrast makes environmental disclosure alignment a suitable case for evaluating whether requirement-level knowledge grounding improves interpretability in multimodal settings and is further illustrated in Table 4, which summarises the key structural differences between GRI 305 and ESRS E1.

Aspect	GRI	ESRS
Section	GRI 305	ESRS E1-4, E1-6
Materiality	Impact Materiality (Voluntary)	Double Materiality (Mandatory under CSRD)
Audience	Broad Stakeholders	Regulators, Investors, Assurance
Targets	Not mandated in GRI 305	Targets required if set; cover Scopes 1–3, 2030/2050 milestones, science-based alignment, etc
Reduction Initiatives	Often Qualitative	Quantified levers required
Biogenic CO2	Varies by reporter	Explicit separation/treatment expected
Assurance Orientation	Low Assurance	Assurance-ready design

Table 4: Comparison of GRI and ESRS Frameworks for GHG Emissions

C Additional Results

C.1 Paragraph-Level Performance under ESRS

Table 5 reports knowledge-augmented alignment performance across individual ESRS requirements. Overall, all three models perform well on requirements centered around explicit quantitative disclosures, particularly P44 (Scope 1–3 and total emissions), P49 (location- and market-based Scope 2 emissions), P50 (disaggregation of Scope 1 and

	Gemini				GPT				Llama			
	Acc	SA	Adj	Opp	Acc	SA	Adj	Opp	Acc	SA	Adj	Opp
E1-4: P33	0.64	0.76	0.25	0.11	0.54	0.71	0.36	0.1	0.61	0.79	0.36	0.03
E1-4: P34	0.46	0.71	0.5	0.04	0.5	0.73	0.46	0.04	0.57	0.77	0.39	0.04
E1-6: P44	0.61	0.80	0.39	0	0.75	0.88	0.25	0	0.93	0.96	0.07	0
E1-6: P46	0.75	0.84	0.18	0.17	0.71	0.86	0.29	0	0.79	0.89	0.21	0
E1-6: P47	0.36	0.57	0.43	0.21	0.36	0.64	0.57	0.07	0.57	0.71	0.29	0.14
E1-6: P48	0.54	0.70	0.32	0.14	0.71	0.82	0.21	0.07	0.71	0.86	0.29	0
E1-6: P49	0.75	0.86	0.25	0	0.82	0.91	0.18	0	0.82	0.89	0.14	0.04
E1-6: P50	0.79	0.89	0.21	0	0.89	0.95	0.11	0	0.64	0.82	0.36	0
E1-6: P51	0.75	0.84	0.18	0.07	0.79	0.88	0.18	0.03	0.79	0.86	0.14	0.07
E1-6: P52	0.57	0.73	0.32	0.11	0.61	0.80	0.39	0	0.54	0.75	0.43	0.04

Table 5: Paragraph-wise evaluation of VLMs on ESRS requirements under the Knowledge-Augmented (KA) setting. Results are reported using Accuracy (Acc), Soft_Acc (SA), Adjacent Error (Adj), and Opposite Error (Opp).

2 emissions), and P51 (Scope 3 category-wise reporting). These requirements achieve the highest accuracies and soft accuracies across models, with LLaMA reaching 0.93 accuracy on P44 and GPT achieving 0.89 accuracy on P50. The consistently low optimism bias for these requirements suggests that VLMs can reliably align disclosures when evidence is presented in structured tables and numerical summaries.

In contrast, all models struggle more with P34 and P47, which require reasoning over methodologies, assumptions, target-setting approaches, reporting-boundary changes, and year-to-year comparability. These requirements exhibit the lowest accuracies across models, with P47 obtaining only 0.36 accuracy for Gemini and GPT. They also produce higher optimism bias and adjacency scores, indicating that models frequently identify partially relevant evidence but have difficulty determining whether all required methodological details are explicitly disclosed.

Comparing models, GPT and LLaMA generally outperform Gemini, although their strengths differ across requirement types. GPT demonstrates the most stable performance across requirements and maintains consistently low optimism bias, while LLaMA achieves the highest scores on several quantitative disclosure requirements, including P44 and P49. However, all models show a similar trend: performance is substantially higher for requirements requiring factual extraction than for those requiring verification of reporting methodology, assumptions, and disclosure completeness. This suggests that requirement-level grounding is effective for identifying explicit evidence but remains challenging for more interpretive and assurance-

oriented ESRS requirements.

C.2 Disclosure-Level Performance under GRI

Table 6 presents requirement-level results for the GRI 305 disclosures under knowledge-augmented prompting. Overall, the models achieve their strongest performance on the core emissions inventory disclosures, particularly 305-1 (Scope 1 emissions) and 305-4 (emissions intensity), where GPT obtains the highest accuracies of 0.75 and 0.75, respectively. All three models also demonstrate low optimism bias for these disclosures, suggesting that explicit and quantitative emissions data can be reliably aligned with GRI requirements.

Performance declines for disclosures that require more detailed contextual information about gases is required. In particular, 305-5 (reduction of GHG emissions) is consistently more challenging across all models.

Among the evaluated models, GPT achieves the most consistent overall performance, leading on five of the seven GRI disclosures. LLaMA performs competitively on 305-5 and maintains the highest soft accuracy for this disclosure (0.82), while Gemini achieves the strongest result on 305-6 (ozone-depleting substances) with an accuracy of 0.72. Across all models, disclosures requiring direct extraction of reported emission values are substantially easier than disclosures involving reduction initiatives, Scope 3 reporting, and pollutant emissions, highlighting that methodological completeness remains the primary challenge in requirement-level GRI alignment.

Figure 5 demonstrates how outputs under the knowledge augmented prompting strategy were classified into Full, Partial, and No Alignment. **Full**

	Gemini				GPT				Llama			
	Acc	SA	Adj	Opp	Acc	SA	Adj	Opp	Acc	SA	Adj	Opp
305-1	0.6	0.8	0.4	0	0.75	0.875	0.25	0	0.72	0.86	0.28	0
305-2	0.4	0.7	0.6	0	0.65	0.825	0.35	0	0.52	0.76	0.48	0
305-3	0.44	0.64	0.4	0.16	0.6	0.75	0.3	0.1	0.48	0.66	0.36	0.12
305-4	0.64	0.8	0.32	0.04	0.75	0.875	0.25	0	0.6	0.8	0.4	0
305-5	0.32	0.6	0.56	0.12	0.55	0.725	0.35	0.1	0.64	0.82	0.36	0
305-6	0.72	0.72	0	0.28	0.65	0.675	0.05	0.3	0.68	0.68	0	0.32
305-7	0.4	0.56	0.32	0.28	0.4	0.575	0.35	0.25	0.28	0.46	0.36	0.36

Table 6: Disclosure-wise evaluation of VLMs on GRI requirements under the Knowledge-Augmented (KA) setting. Results are reported using Accuracy (Acc), Soft_Acc (SA), Adjacent Error (Adj), and Opposite Error (Opp).

Requirement(s)	Output	Alignment
Gases included in the calculation (Scope 1); whether CO ₂ , CH ₄ , N ₂ O, HFCs, PFCs, SF ₆ , NF ₃ , or all	The image states that "In scope 1, we only consider methane and N ₂ O emissions of biogenic emissions, following SBTi guidance, amounting to 2 t of CO ₂ e in 2024.	Full
	The second image explains that GHG emissions are calculated in accordance with the GHG Protocol, which includes various gases.	Partial
	The images do not directly provide the gross Scope 3 emissions. However, the tables on page 11 and 12 list various GHG emissions disclosures, including Scope 1 and Scope 2 emissions, but do not explicitly mention Scope 3 emissions.	No

Figure 5: Example: An alignment based on requirement-level evidence for disclosure 305-1

alignment indicates that all parts of the requirement are explicitly satisfied. **Partial alignment** applies when only some components of the requirement are met or when references are indirect. **No Alignment** is assigned when the response lacks any evidence relevant to the requirement. This categorization ensures fine-grained evaluation of compliance at the requirement level.

Models Internal Knowledge about GRI and ESRS: For GRI, all models generated most of requirements. While Gemini generated 96%, GPT generated 93% and Llama generated 85% of the requirements. Showing higher knowledge corresponding towards GRI. In contrast, for ESRS all models generated weaker paragraphs. Exact numbers can not be calculated for ESRS because of their paragraphs. While qualitative results suggest Gemini appears to possess the most relevant knowledge, followed by GPT. While Llama has the lowest exact knowledge.

Task Complexity: Aligning report content with Sustainability disclosure frameworks is inherently complex. Sustainability reports are typically unstructured, lengthy, and multimodal, containing a mix of narrative text, quantitative tables, charts,

and infographics. Unlike conventional structured benchmarks, the content spans heterogeneous industries and employs diverse terminologies, making it challenging for models to consistently identify relevant disclosures. Further, different frameworks impose requirements at varying levels of granularity and domain specificity. Although many disclosures partially overlap across frameworks, they differ in scope, emphasis, and interpretive expectations, making direct model-driven disclosure alignment a non-trivial interpretability challenge. We analyze alignment behaviour across progressively constrained prompting regimes, examining how varying degrees of guidance and knowledge grounding influence the model’s ability to produce requirement-aware judgments and surface weakly supported or incomplete disclosures. Our objective is to advance AI-assisted, auditable assessment of environmental reporting by clarifying how prompting design influences interpretability and by providing aggregate indicators of reporting coverage introduced in Section 4.

C.3 Baseline Content Understanding

This section provides an example used to verify that the evaluated VLMs can accurately interpret sustainability-report visuals before performing disclosure alignment. The sample includes the original report image (Figure 6) and the corresponding raw output generated by the model (Figure 7). The multi-modality of disclosures can be seen in the report page images. No framework instructions or disclosure requirements were provided; the model was only expected to extract and summarize the information present in the image. This example confirms that the subsequent challenges in disclosure alignment arise from framework-level reasoning rather than limitations in visual or textual comprehension.

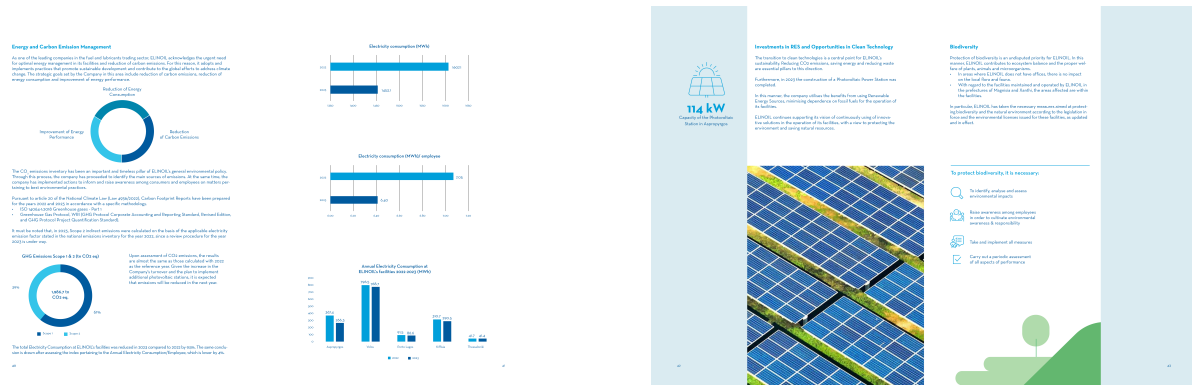


Figure 6: Images from a sustainability report

Prompt:
You are given multiple images from a company's sustainability report. Describe the content provided in the images in detail and image-wise. Be specific with the sections and the related content. Also analyze the infographics, tables, chart, etc. in the images, if present.
Raw Output:
The provided images appear to be from a company's sustainability report, specifically focusing on energy and carbon emission management, investments in renewable energy sources (RES) and clean technology, and biodiversity. Here's a detailed description of the content in each image: **Image 1: Energy and Carbon Emission Management** * The section begins with a description of XYZ's commitment to optimal energy management and reduction of carbon emissions. * A circular diagram illustrates the company's strategic goals: * Reduction of Energy Consumption * Improvement of Energy Performance * Reduction of Carbon Emissions * The CO₂ emissions inventory is a crucial part of XYZ's environmental policy, aiming to identify main emission sources and promote best environmental practices. * A pie chart shows the distribution of GHG Emissions Scope 1 and 2 (tn CO₂ eq): * Scope 1: 39% * Scope 2: 61% * The total electricity consumption at XYZ's facilities decreased by 9.5% in 2023 compared to 2022. * Three bar charts display: * Electricity consumption (MWh) for 2022 and 2023 * 2022: 1607.1 MWh * 2023: 1453.1 MWh * Electricity consumption (MWh)/employee for 2022 and 2023 * 2022: 7.05 MWh/employee * 2023: 6.40 MWh/employee * Annual Electricity Consumption at XYZ's facilities 2022-2023 (MWh) * Aspropyrgos: 367.4 (2022), 266.3 (2023) * Volos: 796.5 (2022), 768.7 (2023) * Porto Lagos: 91.5 (2022), 86.6 (2023) * Kifissia: 310.7 (2022), 290.5 (2023) * Thessaloniki: 41.7 (2022), 41.4 (2023) **Image 2: Investments in RES and Opportunities in Clean Technology and Biodiversity** * The section highlights XYZ's investments in RES and clean technology, including: * A photovoltaic station in Aspropyrgos with a capacity of 114 kW * The construction of a Photovoltaic Power Station in 2023 * The company's vision for using innovative solutions to protect the environment and save natural resources is emphasized. * A section on biodiversity discusses XYZ's efforts to: * Protect ecosystem balance and welfare of plants, animals, and microorganisms * Identify, analyze, and assess environmental impacts * Raise awareness among employees to cultivate environmental awareness and responsibility * Take and implement measures to protect biodiversity * Carry out periodic assessments of performance The infographics, tables, and charts in the images effectively illustrate XYZ's progress and goals in energy and carbon emission management, investments in RES and clean technology, and biodiversity conservation. The data presented demonstrates the company's commitment to sustainability and environmental responsibility.

Figure 7: Sample prompt used with model output

C.4 GRI vs ESRS Findings and Compliance Critique

For the same set of reports, coverage breadth is steadier and generally higher under GRI because

core requirements are widely disclosed, whereas ESRS breadth is lower and more variable because coverage depends on paragraph-level evidence that many sectors only partially provide.

Gap	Sectors										
	Consumer Goods	Extractives & Mineral Processing	Financials	Food & Beverage	Health Care	Infrastructure	Renewable Resources & Alternative Energy	Resource Transformation	Services	Technology & Communications	Transportation
Scope 3 emissions incomplete/unclear	☒	☒	☐	☐	☒	☐	☐	☐	☐	☒	☐
Biogenic CO ₂ missing	☒	☒	☒	☒	☒	☐	☐	☒	☒	☒	☒
Reduction Initiatives lack quantification	☐	☐	☐	☒	☐	☐	☐	☒	☒	☐	☒
Consolidation approach unclear	☒	☒	☒	☒	☐	☒	☒	☐	☒	☒	☐
Pollutant Data Missing	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒

(a) GRI 305

Gap	Sectors										
	Consumer Goods	Extractives & Mineral Processing	Financials	Food & Beverage	Health Care	Infrastructure	Renewable Resources & Alternative Energy	Resource Transformation	Services	Technology & Communications	Transportation
Scope 3 not disclosed by each significant category	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒
Scope 1/2 not disaggregated	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒
Associates/JVs/unconsolidated entities not included by operational control	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒
Scope 1: % of emissions under ETS not disclosed	☒	☒	☐	☒	☐	☒	☐	☒	☒	☐	☐
Scope 2: location-based and market-based emissions both not provided	☒	☒	☐	☒	☒	☐	☐	☒	☒	☒	☐
Total GHG not disclosed as dual totals	☒	☒	☐	☒	☒	☒	☐	☒	☒	☒	☐
Boundary changes not explained	☒	☒	☒	☒	☒	☒	☐	☒	☒	☒	☐
Reduction targets are set not disclosed / unclear	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒
Targets lack absolute/intensity basis or	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒
Targets missing required milestones (≥2030; if available 2050)	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒
Target methodology & critical assumptions	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒
Decarbonisation levers not described with quantitative contributions	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒

(b) ESRS E1-4, E1-6

Figure 8: Sector-wise gaps across reporting framework disclosures

The common gaps (Figure 8) that appear repeatedly across sectors are *Significant Scope 3 details*, *boundary/operational-control clarity*, *Location-based/Market-based handling with dual totals*, *Emissions Trading System share*, and *auditable target mechanics*. These are precisely those that ESRS requires to be explicit, and their recurrence explains the higher ESRS gap density seen in Figure 4. Practically, the sector path to improvement is to raise ESRS coverage while shrinking those recurring gaps, so that breadth converts into assurance-ready depth rather than remaining as partial alignment with persistent methodological deficits.

C.5 Prompt Templates

For reproducibility, we provide the prompts used (Figure 9, 10, 11) under the three conditional settings: Open-Ended (OE), Framework-Specific (FrS), and Knowledge-Augmented (KA). Deterministic decoding was employed to ensure reproducibility by eliminating stochasticity during text generation and consistently selecting the most probable token at each decoding step.

Open Ended Prompt
You are a sustainability reporting analyst. Given one or more images from a company's sustainability report, identify whether the content aligns with GRI and/or ESRS emission-related disclosures based on explicit or implicit evidence in the report, disclosures, and reporting requirements reflected in the content. For GRI, provide the disclosure-wise analysis while for ESRS, provide the paragraph-wise alignment. For each identified disclosure or requirement: • Provide the disclosure/requirement code and title. • Determine whether it is Fully Aligned, Partially Aligned, or Not Aligned. For partially aligned or non-aligned disclosures, identify the missing information required for full compliance. Consider all available report content, including text, tables, charts, and visual elements, and use official disclosure or requirement identifiers whenever possible.

Figure 9: Prompt used for Open-ended setting

Framework Specific Prompt
You are a sustainability reporting analyst. Given one or more images from a company's sustainability report, assess alignment with the specified sustainability reporting framework (GRI/ESRS). Instructions: 1. Identify the emission-related disclosures or requirements addressed in the images. 2. For each disclosure/requirement: • Provide the official disclosure/requirement code and title. • Determine whether it is Fully Aligned, Partially Aligned, or Not Aligned. • Cite the supporting image(s) and briefly justify the assessment. 3. Identify any missing information required for disclosures or requirements that are partially or not aligned. 4. Provide an overall alignment assessment for the framework. Output Format: Disclosure/Requirement: <code> - <title> • Alignment: • Supporting Image(s): • Rationale: Overall Alignment: Missing Elements:

Figure 10: Prompt used for Framework-specific setting

Knowledge Augmented Prompt
You are a sustainability reporting analyst. You are provided with one or more images from a company's sustainability report and a disclosure along with its reporting requirements for GRI (or ESRS) framework. Disclosure: {disclosure} Disclosure Name: {disclosure_name} Requirements: {requirements} Assess how well the disclosure requirements are satisfied based on all provided images. Instructions: 1. Evaluate each requirement independently. 2. For each requirement, determine whether it is Fully Addressed, Partially Addressed or Not Addressed 3. Based on all requirement-level assessments, determine the overall disclosure alignment as Fully Aligned, Partially Aligned or Not Aligned Output Format: Req 1 - <requirement description> - <alignment status> Req 2 - <requirement description> - <alignment status>... Overall Alignment - <overall alignment status> Return only the required output in the specified format. Do not provide explanations, reasoning, or additional information.

Figure 11: Prompt used for Knowledge-Augmented setting