Exploring Multimodal Language Models for Sustainability Disclosure Extraction: A Comparative Study

Tanay Kumar Gupta
TCS Research
New Delhi, India
gupta.tanay@tcs.com

Tushar Goel
TCS Research
New Delhi, India
t.goel@tcs.com

Ishan Verma
TCS Research
New Delhi, India
ishan.verma@tcs.com

Abstract

Sustainability metrics have increasingly become a crucial non-financial criterion in investment decision-making. Organizations worldwide are recognizing the importance of sustainability and are proactively highlighting their efforts through specialized sustainability reports. Unlike traditional annual reports, these sustainability disclosures are typically text-heavy and are often expressed as infographics, complex tables, and charts. The non-machine-readable nature of these reports presents a significant challenge for efficient information extraction. The rapid advancement of Vision Language Models (VLMs) has raised the question whether these VLMs can address such challenges in domain specific task. In this study, we demonstrate the application of VLMs for extracting sustainability information from dedicated sustainability reports. Our experiments highlight the limitations in the performance of several open-source VLMs in extracting information about sustainability disclosures from different type of pages.

1 Introduction

In recent years, we have witnessed a significant growth in inclusion of non-financial factors particularly sustainability in corporate reporting. As per KPMG's recent sustainability reporting survey[1], reporting on sustainability has become part of business as usual for 96% of the world's largest 250 companies and a majority of the top 100 companies in each country. Driven by continued stakeholder demand for transparency and consistency in sustainability data disclosures, several standards have been proposed to harmonize sustainability reporting. Frameworks like Global Reporting Initiative (GRI), Sustainability Accounting Standards Board (SASB), and UN Sustainable Development Goals (SDG), have been developed to streamline the reporting around various sustainability indicators (Chofreh and Goni, 2017). Despite being part of business-critical disclosures, sustainability reporting remains mostly unstructured, unlike the standardized annual financial reports. With no common reporting template, these reports lack consistency, relying mostly on charts, infographics and text, and are generally published in non-machine-readable PDF formats (Ruggiero and Bachiller, 2023). Extracting relevant information from these unstructured reports takes several person hours of efforts and is prone to mistakes and influence by personal judgment. Hence, automation of sustainability information extraction from reports can reduce processing time and let stakeholders focus more on decision making. In this work, we have used the GRI framework as reference. The GRI framework offers sustainability standards in three categories: Economy (GRI 200), Environment (GRI 300), and Social (GRI 400), for further details refer A.1. Organisations use various indicators listed under these three to report their sustainability activities.

Recent advancements in natural language processing with availability of Large Language Models (LLM) and Vision Language Models (VLM) viz Llama (Touvron et al., 2023), Gemini (Team et al., 2023, 2024), Phi (Abdin et al., 2024), ChatGPT (Achiam et al., 2023), LLaVa (Liu et al., 2023a,b), etc have opened a new dimension to multi-modal information extraction. Significant number of efforts have been made to utilize LLMs for extracting information from sustainability or Environmental, Social, and Governance (ESG) reports. For instance, ClimateBERT (Webersinke et al., 2021) is a transformer model fine-tuned for climate-related classification tasks, ChatReport (Ni et al., 2023) is an LLM-based tool that evaluates companies' sustainability reports according to the TCFD guidelines. ESGReveal (Zou et al., 2023) is a tripartite framework leveraging LLMs and RAG to extract and analyze ESG data, offering benchmarks for corporate reporting. DocQA (Mishra et al., 2024)

[1] https://kpmg.com/xx/en/our-insights/esg/the-move-to-mandatory-reporting.html

is a platform for question answering over sustainability reports using RAG framework. Bronzini et al. (Bronzini et al., 2024) used LLMs to construct knowledge graphs for analyzing ESG disclosures from sustainability reports. However, to the best of our knowledge, the use of VLMs for ESG data extraction is yet to be explored. Also, LLMs still face challenges in effectively handling domain-specific tasks in a zero-shot setting(Yao et al., 2023). In this work, we have evaluated the performance of open source VLMs on the task of extracting sustainability disclosures from sustainability reports. We highlight the shortcomings of these models on the extraction task.

2 Data Curation & System Architecture

We collected around 700 English language sustainability reports from the SASB website[2]. Many organizations include a GRI index table in their sustainability reports which lists GRI disclosures with their reference, such as page numbers, links, and section headers. The format varies across reports, as can be seen in A.2. We used GRI index tables from these reports to generate the benchmark evaluation dataset, focusing on those with internal references only for listed GRI disclosures. We applied a keyword driven heuristic-based filtering method to identify reports containing a GRI Index table, specifically checking if the index mention appeared in the table of contents. This filtering process left us with 380 reports. Next, we manually annotated the page ranges in the reports where the GRI Index tables appeared. These tables were extracted using a combination of Table Transformer (Smock et al., 2022) and Llama 3 (Dubey et al., 2024) with human in the loop in <GRI disclosure, Page Number> format. Comparison of our approach with VLM based table extractions are shown in appendix A.2.

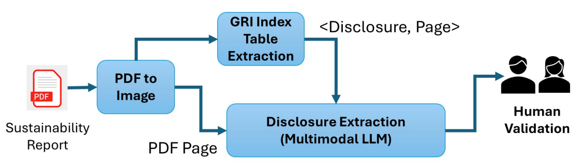


Figure 1: System Flow Diagram

Figure 1 shows the process flow followed for the work. After identifying the relevant page image and corresponding GRI disclosure for which data needs to be extracted, we provided the page image along with a contextual prompt to the VLMs. The prompt was designed to guide the model in extracting specific information based on the GRI disclosure. The VLM output is then validated by human experts. The next section outlines the VLMs used in our experiments, a brief overview of their architecture and selection rationale.

3 Experiments

We experimented with the following VLMs for extracting information related to GRI disclosures from the report pages:

- **Llama 3.2 Vision** - Llama-3.2-11B-Vision-Instruct[3] is an instruction fine-tuned model of Llama 3.2 vision (Dubey et al., 2024) which integrates image encoder followed by image adapter and language model decoder.
- **Qwen2 VL** - Qwen2-VL-7B-Instruct[4] is an instruction fine-tuned model of Qwen2 (Wang et al., 2024) which integrates 675M parameter vision encoder with 7.6B parameter language model decoder. It uses naive dynamic resolution to process any resolution image along with multi-modal rotary position embedding to extrapolate longer sequences.
- **LLaVA** - LLaVA-v1.6-mistral-7b-hf [5] is an instruction fine-tuned model of LLaVA-NeXT (Liu et al., 2024) which combines a vision encoder and a connector to connect with mistral (Jiang et al., 2023) LLM for joint vision-language tasks.

These models were selected due to their proven accuracy on similar datasets, such as DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), InfographicVQA (Mathew et al., 2022), and MMMU (Yue et al., 2024). Additionally, they represent some of the recent advancements in the field, ensuring that the models used are both relevant and capable of handling the complexities of the task at hand. In the later section, we will discuss about the short comings of these models in specific settings.

4 Results & Analysis

We conducted experiments to validate the extractions of 74 unique GRI disclosures across 10 sus-

[2]https://sasb.ifrs.org/company-use/sasb-reporters/

[3]https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

[4]https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

[5]https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

tainability reports, with a total of 640 pages analyzed. The results were assessed by two independent domain experts, with a Cohen's Kappa score(Cohen, 1960) of 0.9, indicating strong consistency in the evaluation process. Partial extractions were considered incorrect during validation. The requirement for manual validation of the outputs limited the number of reports included in the experiments. These pages included data presented as text, tables, charts, infographics, or a combination of these formats (Examples shown in Appendix A Figure 2). Around 60% of the pages have data in more than one format. In Table 1, we present the distribution of pages as per the data formats.

Page Type	Count	Page Type	Count
text	125	**table+infographics**	10
table	125	**chart+infographics**	7
infographics	5	**text+chart**	90
text+infographics	163	**text+chart+table**	46
text+table	39	**text+chart+infographics**	30

Table 1: Distribution of Pages as per Data format

The prevalence of "*text*" and "*text + infographics*" highlights the reliance on textual descriptions and visual information in sustainability reporting. However, the reports also incorporate a substantial number of tables and combined modalities like "*text + chart*" and "*text + chart + table*", emphasizing the multimodal nature of these documents. The low incidence of standalone "*infographics*" and "*chart + infographics*" likely stems from the common practice of including accompanying text within these visuals, resulting in a higher prevalence of combined "*text + infographics*" entries.

We expand our analysis by including key metrics that evaluate the effectiveness of VLMs in GRI disclosure extraction. For validation purposes, we considered three distinct cases:

- **Correct**: The model correctly extracts the relevant information from the page.
- **Incorrect**: The model extracts information from the page; however, it fails to provide the correct answer due to misinterpretation, incomplete understanding, or inaccurate reasoning. This could involve selecting the wrong data point from a chart/table or misinterpreting a statement in the text.
- **Hallucination**: The model provides information that is not present on the page. This indicates that the response isn't properly based on the page content, and the model is generating the response based on its training data.

The accuracy of information extraction using VLMs is calculated as the percentage of correct, incorrect, or hallucinated instances, out of the total number of instances for a given GRI category or page type. Table 2 shows VLMs accuracy in extracting information across GRI categories. The results reveal varying performance across these categories for all the models, indicating sensitivity to the specific content and language used within each. For instance, while Llama 3.2 vision demonstrates a relatively high accuracy in the Economic category, likely due to the structured and quantifiable nature of economic disclosures, its performance dips in the Environment and Social categories. As for LLaVa and Qwen2, a striking negative result is the substantial proportion of "incorrect" predictions, often exceeding "correct" ones. This highlights models issue with fine-grained comprehension and information localization within the document. Furthermore, the presence of hallucination raises concerns about reliability, specially for LLaVa in the Environment category. This is due to LLaVa's tendency to generate descriptive summaries of page images, rather than focusing on precise information extraction, which might contributed to the fabrication of information. Qwen2 displays a different pattern, exhibiting a lower hallucination rate but a high proportion of incorrect predictions. This behavior suggests a potential weakness in Qwen2's ability to perform fine-grained information extraction. Our observations indicate that Qwen2 occasionally provides section headlines or the names of GRI disclosures themselves as answers, even when more specific information is requested in the prompt. This tendency to offer labels rather than detailed content contributes to the increased rate of incorrect predictions. The impact of different page types on the VLM's accuracy is demonstrated in Table 3, revealing a significant performance gap based on content and layout complexity. While text-heavy pages achieve moderate accuracy for Llama 3.2 and Qwen2, LLaVa struggles, suggesting differences in handling textual information. Complex layouts involving tables, charts, or their combinations like "*text + table*", "*chart + infographics*" present consistent challenges for all VLMs, with accuracy often falling below 50%, highlighting difficulties in deciphering information embedded within structured or visually formatted elements. The inclusion

Category	Llama 3.2 Vision Instruct			LLaVa v1.6			Qwen2 VL Chat		
	Correct	Incorrect	Hallucination	Correct	Incorrect	Hallucination	Correct	Incorrect	Hallucination
Economic	68.42	28.95	2.63	50	36.84	13.16	39.47	60.53	0
Environmental	50.47	45.04	4.47	38.02	40.25	21.72	45.36	51.43	3.19
Social	62.15	35.45	2.39	36.65	44.22	19.12	57.37	41.43	1.19

Table 2: Accuracy(%) of VLMs in extracting information across GRI categories

Page Type	Llama 3.2 Vision Instruct			LLaVa v1.6			Qwen2 VL Chat		
	Correct	Incorrect	Hallucination	Correct	Incorrect	Hallucination	Correct	Incorrect	Hallucination
text	66.4	32	1.6	38.4	41.6	20	55.2	44	0.8
table	47.2	44	8.8	40.8	40.8	18.4	36	57.6	6.4
infographics	80	20	0	40	60	0	100	0	0
text+infographics	55.83	41.72	2.45	30.06	44.78	25.15	43.56	55.83	0.61
text+table	41.03	58.97	0	46.15	25.64	28.21	56.41	41.03	2.56
table+infographics	100	0	0	0	0	100	80	20	0
chart+infographics	14.29	85.71	0	14.29	57.14	28.57	42.86	57.14	0
text+chart	66.67	31.11	2.22	51.11	46.67	2.22	64.44	35.56	0
text+chart+table	47.83	52.17	0	56.52	39.13	4.35	43.49	52.17	4.35
text+chart+infographics	66.67	23.33	10	26.67	40	33.33	50	50	0
Overall	57.19	39.38	3.44	38.91	41.41	19.69	49.38	48.59	2.03
No Data	22.11	71.63	6.25	48.55	41.34	10.09	40.38	57.69	1.92

Table 3: Accuracy(%) of VLMs in information extraction from different page types

of "*infographics*" seems to have a varied impact. Llama and Qwen2 achieve high accuracy with standalone infographics, but performance decreases when combined with text, indicating challenges in integrating multimodal information. Hallucination tendencies, particularly prominent in LLaVa for heavily structured pages like "*text + chart + infographics*", "*chart + infographics*", suggest a potential link between difficulty in processing specific page structures and the tendency to hallucinate. Overall, Llama3.2 and Qwen2 exhibit comparable performance, while LLaVa lags, underscoring the need for further research into how VLMs process diverse page elements and mitigate hallucination, especially in complex layouts. Few examples for the same are shown in appendix A.3.

The "No Data" case, shown in the last row of Table 3 presents a notable finding from our experiments. In 208 instances, the relevant data for disclosures listed in the GRI index table was not present on the referenced pages. We then assessed the performance of VLMs on pages where GRI disclosure information is expected but absent. This scenario tested the models' ability to handle missing data and avoid generating potentially misleading or fabricated responses. The results reveal a concerning trend: Llama and Qwen2 exhibits a high proportion of "incorrect" predictions in this context whereas LLaVa achieves highest accuracy in saying "no, the information is not present". Llama, while demonstrating the lowest percentage of correct responses, struggles significantly, with over 72% of its predictions classified as incorrect. Interestingly, while the hallucination rates are low for all models, the substantial proportion of incorrect predictions suggests that the models may be attempting to answer by relying on contextual clues or related information, even when the specific data point is missing. This highlights a critical limitation: the models appear unable to reliably identify and flag the absence of required information, instead attempting to provide an answer, even if it is incorrect. This behavior underscores the need for improved mechanisms to detect and handle missing data.

5 Conclusion

In this work, we explored the feasibility of using Vision-Language Models (VLMs) for sustainability data extraction from multimodal PDF page images. Our experiments concluded that no single VLM can efficiently manage all data formats. We found that Llama performs best on text-based pages but is prone to incorrect responses. The LLaVa model frequently experiences hallucinations, while Qwen exhibits similar accuracy for both correct and incorrect responses. This study opens potential future research directions, such as integrating model strengths, fine-tuning for improved performance, and using knowledge-infused prompts for better extraction. It is also important to address cases with no data, focusing on extending VLMs ability to recognize and respond to information gaps.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

Abdoulmohammad Gholamzadeh Chofreh and Feybi Ariani Goni. 2017. Review of frameworks for sustainability implementation. *Sustainable Development*, 25(3):180–188.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Lokesh Mishra, Cesar Berrospi, Kasper Dinkla, Diego Antognini, Francesco Fusco, Benedikt Bothur, Maksym Lysak, Nikolaos Livathinos, Ahmed Nassar, Panagiotis Vagenas, et al. 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 Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, et al. 2023. Chatreport: Democratizing sustainability disclosure analysis through llm-based tools. *arXiv preprint arXiv:2307.15770*.

Pasquale Ruggiero and Patricia Bachiller. 2023. Seeing more than reading: The visual mode in utilities' sustainability reports. *Utilities Policy*, 83:101610.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Jing Yao, Wei Xu, Jianxun Lian, Xiting Wang, Xiaoyuan Yi, and Xing Xie. 2023. Knowledge plugins: Enhancing large language models for domain-specific recommendations. *arXiv preprint arXiv:2311.10779*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yi Zou, Mengying Shi, Zhongjie Chen, Zhu Deng, ZongXiong Lei, Zihan Zeng, Shiming Yang, HongXiang Tong, Lei Xiao, and Wenwen Zhou. 2023. Esgreveal: An llm-based approach for extracting structured data from esg reports. *arXiv preprint arXiv:2312.17264*.

A Appendix

Figure 2 illustrates several examples of pages from sustainability reports that exhibit multimodal content. These examples highlight the complexity involved in processing such pages, where visual elements such as text, tables, charts, infographics and their combinations are present.

A.1 GRI Framework

The Global Reporting Initiative (GRI)[6] framework is a globally recognized standard for sustainability reporting that helps organizations disclose their environmental, social, and economic impacts. It provides structured guidelines to ensure transparency, consistency, and comparability in sustainability reports. The GRI framework consists of several series, each addressing specific areas of sustainability:

- **200 Series** focuses on economic factors, guiding organizations to report on their Economic performance(201), Market presence(202), Anti-competitive behaviour(206) etc.

- **300 Series** deals with environmental aspects, covering topics such as Energy(302), Water and Effluents(303), Emissions(305), and biodiversity(304) etc.

- **400 Series** addresses social factors, including labor practices such as Child labor(408), Forced or Compulsory labor(409), Training and Education(404), Occupational health and safety(403) etc.

Table 4 provides a sample breakdown of the GRI series, illustrating how each series is further subdivided into specific disclosures.

[6]https://www.globalreporting.org/

Economic (200)	**Anti-corruption (205)**	**Operations Assessed for Risks related to Corruption(205-1)**
Environment (300)	**Energy (302)**	**Energy consumption within the organization (302-1)**
		Energy intensity (302-3)
	Water (303)	**Water Withdrawal (303-3)**
Social (400)	**Occupational Health & Safety (403)**	**Promotion of Worker Health (403-6)**

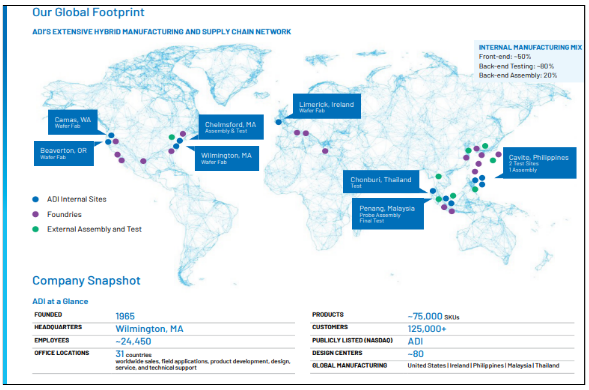
Table 4: GRI hierarchy example

A.2 GRI Index Table

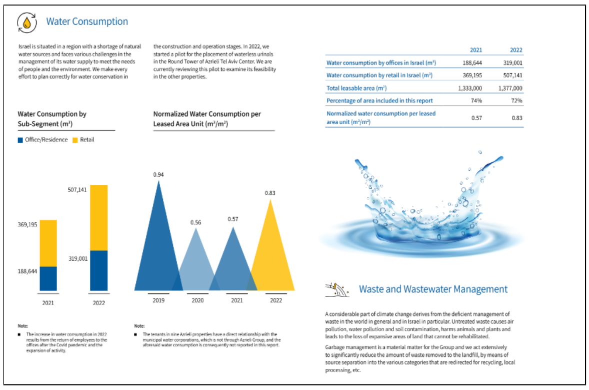
In Figure 3, we have shown few examples of GRI index tables. These tables have information about GRI disclosures along their data references in the reports. These references can be present in form of page number along with or without textual information, internal and external links, name of other reports, and section headers within the report along with other metadata. These formats differ from one report to another. To extract the information about GRI disclosures mentioned in the table, it is required to have the GRI disclosure data in <GRI Disclosure, Page Number> format. While validating, we have relaxed the assumption of strict page number such as if we get p.38, we mark it correct. We have validated it for 10 sustainability reports. Table 5 shows the accuracy of VLMs for GRI index table extraction and our approach Table Transformer + Llama 3. Here, Llama 3.2 Vision achieved 83.5% accuracy but struggled with providing direct page numbers when text is involved. Additionally, it had difficulty in maintaining format consistency. In contrast, Qwen2 faced issues in detection of GRI disclosure and performs poorly when multiple tables were present and achieved only 53.5% accuracy. LLaVa, on the other hand, consistently produced repetitive results, falling short of expectations. In comparison, our approach delivers accurate page numbers and achieves superior accuracy of 93%, making it the preferred method for GRI Index extraction. This observation highlights the limitations of Qwen2 and LLaVa in accurately extracting information from page images containing large tables, a scenario that is not commonly encountered.

	TableTransformer + Llama 3 Instruct	**Llama 3.2 Vision**	**Qwen2 VL Chat**
Accuracy	**93%**	83.5%	53.5%

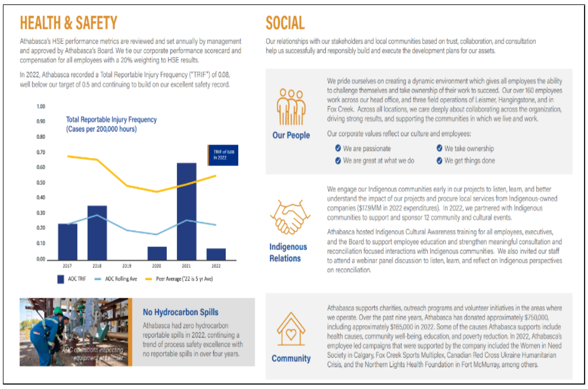
Table 5: Accuracy of different models in GRI Index Table Extraction



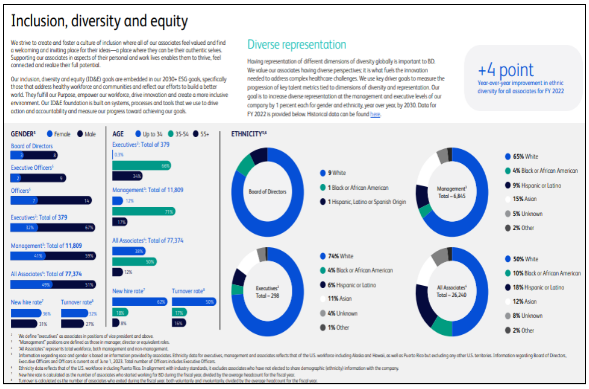
Page type: Table+Infographics



Page type: Text+Chart+Table

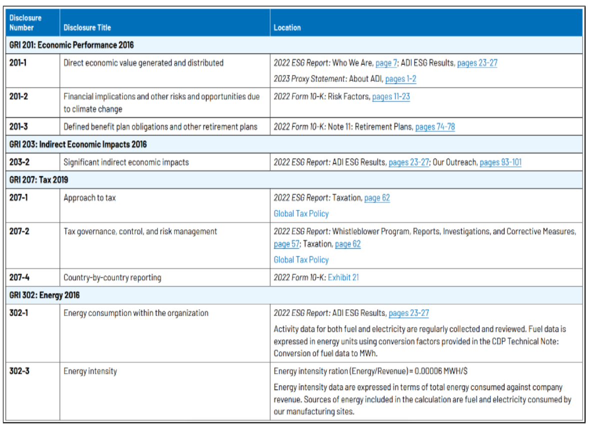


Page type: Text+Chart+Infographics

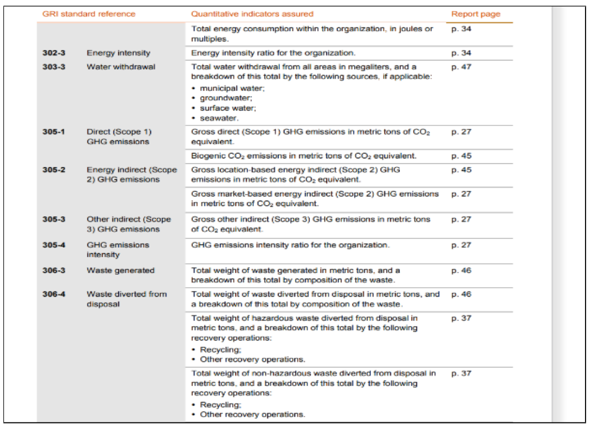


Page type: Text+Chart

Figure 2: Examples of pages from sustainability reports exhibiting multimodal content, highlighting the complexity involved in processing visual elements to extract the relevant data.



Disclosure Number	Disclosure Title	Location
GRI 201: Economic Performance 2016		
201-1	Direct economic value generated and distributed	*2022 ESG Report:* Who We Are, page 7; ADI ESG Results, pages 23-27 *2023 Proxy Statement:* About ADI, pages 1-2
201-2	Financial implications and other risks and opportunities due to climate change	*2022 Form 10-K:* Risk Factors, pages 11-23
201-3	Defined benefit plan obligations and other retirement plans	*2022 Form 10-K:* Note 11: Retirement Plans, pages 74-78
GRI 203: Indirect Economic Impacts 2016		
203-2	Significant indirect economic impacts	*2022 ESG Report:* ADI ESG Results, pages 23-27; Our Outreach, pages 93-101
GRI 207: Tax 2019		
207-1	Approach to tax	*2022 ESG Report:* Taxation, page 62 Global Tax Policy
207-2	Tax governance, control, and risk management	*2022 ESG Report:* Whistleblower Program, Reports, Investigations, and Corrective Measures, page 57; Taxation, page 62 Global Tax Policy
207-4	Country-by-country reporting	*2022 Form 10-K:* Exhibit 21
GRI 302: Energy 2016		
302-1	Energy consumption within the organization	*2022 ESG Report:* ADI ESG Results, pages 23-27 Activity data for both fuel and electricity are regularly collected and reviewed. Fuel data is expressed in energy units using conversion factors provided in the CDP Technical Note: Conversion of fuel data to MWh.
302-3	Energy intensity	Energy intensity ration (Energy/Revenue) = 0.00006 MWH/$ Energy intensity data are expressed in terms of total energy consumed against company revenue. Sources of energy included in the calculation are fuel and electricity consumed by our manufacturing sites.



GRI standard reference		Quantitative indicators assured	Report page
		Total energy consumption within the organization, in joules or multiples.	p. 34
302-3	Energy intensity	Energy intensity ratio for the organization.	p. 34
303-3	Water withdrawal	Total water withdrawal from all areas in megaliters, and a breakdown of this total by the following sources, if applicable: • municipal water; • groundwater; • surface water; • seawater.	p. 47
305-1	Direct (Scope 1) GHG emissions	Gross direct (Scope 1) GHG emissions in metric tons of CO_2 equivalent.	p. 27
		Biogenic CO_2 emissions in metric tons of CO_2 equivalent.	p. 45
305-2	Energy indirect (Scope 2) GHG emissions	Gross location-based energy indirect (Scope 2) GHG emissions in metric tons of CO_2 equivalent.	p. 45
		Gross market-based energy indirect (Scope 2) GHG emissions in metric tons of CO_2 equivalent.	p. 27
305-3	Other indirect (Scope 3) GHG emissions	Gross other indirect (Scope 3) GHG emissions in metric tons of CO_2 equivalent.	p. 27
305-4	GHG emissions intensity	GHG emissions intensity ratio for the organization.	p. 27
306-3	Waste generated	Total weight of waste generated in metric tons, and a breakdown of this total by composition of the waste.	p. 46
306-4	Waste diverted from disposal	Total weight of waste diverted from disposal in metric tons, and a breakdown of this total by composition of the waste.	p. 46
		Total weight of hazardous waste diverted from disposal in metric tons, and a breakdown of this total by the following recovery operations: • Recycling; • Other recovery operations.	p. 37
		Total weight of non-hazardous waste diverted from disposal in metric tons, and a breakdown of this total by the following recovery operations: • Recycling; • Other recovery operations.	p. 37

Figure 3: Examples of GRI Index Tables

Figure 4: Page images corresponding to Table 6 for data extraction

GRI Disclosure	Page type	VLM Model	Information Extracted using VLM	Validation
Water Consumption (303-5)	Table (Image A)	LLaVa	"Total Water Consumption" with a value of 807 million m3. "Total Water Consumption from areas with high stress" with a value of 807 million m3 in the same table.	Hallucination
Average hours of training per year per employee (404-1)	Text + Infographics (Image B)	Qwen	['8,000']	Incorrect
Nitrogen oxides (NOX), sulfur oxides (SOX), and other significant air emissions (305-7)	Table (Image C)	Llama	**Nitrogen Oxides (NOX)** * **2019**: 0 * *2020**: 0 * **2021**: 0 * **2022**: 0 *Sulfur Oxides (SOX)** * **2019**: 0 * *2020**: 0 * **2021**: 0 * **2022**: 0	Hallucination
Confirmed incidents of corruption and actions taken (205-3)	Text (Image D)	Qwen	['Confirmed incidents of corruption and action taken.']	Incorrect
GHG emissions intensity (305-4)	Text+Chart +Table (Image E)	Llama	Total Normalized Scope 1+Scope 2 Greenhouse Gas Emissions per Area Unit (tCO2e/m2). Scope 1 + Scope 2 Greenhouse Gas Emissions (tCO2e). These are the only references to GHG emissions intensity in the image.	Incorrect
Waste diverted from disposal (306-4)	Text+Chart+ Infographics (Image F)	LLaVa	As part of our ongoing commitment to sustainable business practices, we continuously monitor and improve our waste management processes. By implementing new technologies and strategies, we aim to achieve zero landfill waste generation across all sites by 2030. This initiative aligns with our broader goal of achieving net zero emissions by 2050.	Hallucination
Reduction of GHG emissions (305-5)	Text+ Infographics (Image G)	Qwen	"FY 222 GHG Emissions, by scope and category 7,587,446 metric tonnes CO2e SCOPE 1 & 2 Direct and indirect emissions Total: 6% SCOPE 3 CATEGORY 1 Purchased goods and services Total: 42% SCOPE 3 CATEGORY 4 Upstream transportation and distribution Total: 8% SCOPE 3 CATEGORY 9 Downstream transportation and distribution Total: 4% SCOPE 3 CATEGORY 11 ....."	Correct

Table 6: Examples of data extraction from images using different VLMs along with its validation

A.3 Examples of GRI Disclosure Extraction

Figure 4 shows few reference page images. Information extracted through different VLMs for few GRI disclosures from these images are shown in Table 6. The analysis of these examples are as follows:

- **Row 1 - Image A**: An attempt was made to extract information related to water consumption; however, LLaVa failed to correctly retrieve the information. The extracted value did not correspond to any data present on the page, resulting in a case of hallucination.
- **Row 2 - Image B**: Qwen2 was tasked to extract the information about Average hours of training per year per employee but it erroneously extracted number of people managers benefited from the manager curriculum which is incorrect.
- **Row 3 - Image C**: It shows another case of hallucination where Llama 3.2 generated false value "0" for the disclosure Nitrogen oxides (NOX), sulfur oxides (SOX), and other significant emissions. This example also highlight the case of "**No Data**" as there is no information available corresponding to the GRI disclosure "Nitrogen oxides (NOx), sulfur oxides (SOx), and other significant air emissions(305-7)" on the page.
- **Row 4 - Image D**: Qwen2 failed to extract any information from this image, which contained only textual data.
- **Row 5 - Image E**: Llama 3.2 provided the title of charts as an answer which is an incorrect response.
- **Row 6 - Image F**: In this complex image which contains text, chart and infographics all together, LLaVa exhibited hallucination by providing a statement "we aim to achieve zero landfill waste generation across all sites by 2030." as an answer. This information is not present on the page.
- **Row 7 - Image G**: Although complex data representation in their in the page image, Qwen2 is able to extract correct information from the infographics.