ClariESG: An End-to-End System for ESG Analysis over Complex Tables in Corporate Reports

Marta Santacroce, Michele Luca Contalbo, Sara Pederzoli, Riccardo Benassi, Valeria Venturelli, Matteo Paganelli, Francesco Guerra

University of Modena and Reggio Emilia, Modena, Italy,
{name.surname}@unimore.it

Abstract

Sustainability reports contain rich Environmental, Social and Governance (ESG) information, but their heterogeneous layouts and complex multi-table structures pose major challenges for LLMs, especially for unit normalization, cross-document reasoning, and precise numerical computation. We present ClariESG, an end-to-end system that couples robust table extraction with a structured prompting framework for multi-table filtering, normalization, and program-of-thought reasoning. On ESG-focused multi-table benchmarks, ClariESG consistently outperforms standard prompting and provides transparent, auditable reasoning, supporting more reliable ESG analysis and greenwashing detection in real-world settings.

1 Introduction

Companies increasingly publish *sustainability reports*, i.e., documents that describe their ESG performance, policies, and non-financial impacts, to comply with sustainability standards such as GRI (GRI, 2024). These reports are a key source of ESG data, used in sustainable finance, corporate accountability, and policy evaluation. In Europe alone, assets managed under responsible investment principles reached approximately €6.6 trillion in 2024, representing nearly 38% of total managed assets (Heflich and Saulnier, 2024).

To fully exploit the analytical potential of these disclosures, the European Union is introducing the *European Single Access Point* (CSR, 2022), a unified platform that will begin collecting sustainability reports starting in 2026. While this centralization will enable unprecedented large-scale, data-driven ESG analysis, it also underscores the need for automated methods capable of consistently interpreting complex and heterogeneous disclosures in a transparent and explainable manner.

Recent advances in Large Language Models (LLMs) offer a promising direction for automating ESG knowledge extraction, thanks to their ability to interpret unstructured and heterogeneous data. Nevertheless, sustainability reports remain highly complex documents that challenge even state-of-the-art models. Evidence of this emerges from the GRI-QA benchmark (Contalbo et al., 2025), a dataset for single- and multi-table question answering over sustainability reports. When tested on multi-table scenarios, GPT-based models show limited performance even with clean, expert-curated tables, indicating that the difficulty lies not in OCR noise or table detection errors but in the intrinsic reasoning demands of the domain. This empirical analysis highlights three main challenges.

First, the tabular data in sustainability reports frequently exhibits non-standard structures, including hierarchical layouts, merged headers, and company-specific schemas. Second, the language used in these documents is highly domain-specific, combining technical terminology with performance indicators that vary in definition, scope, and units of measure. Finally, interpreting the disclosed information can require complex numerical reasoning, involving the combination and comparison of indicators distributed across multiple tables, sometimes even in different documents.

The literature has proposed several end-to-end systems for processing and querying sustainability reports with LLMs (Zou et al., 2023; Ni et al., 2023; Vaghefi et al., 2023; Singh et al., 2024; Wrzalik et al., 2024; Nguyen et al., 2025; Hsu et al., 2024); however, these systems generally do not provide explicit support for tabular content or for complex quantitative reasoning, particularly in scenarios requiring cross-table numerical aggregation and interpretation.

To address these limitations, we present ClariESG[1], an end-to-end LLM-based system designed

[1] The code and demonstration video are available at `https://github.com/softlab-unimore/ClariESG.git` and `https://youtu.be/noQ-5ya6cOE`

to support automated analysis of corporate sustainability reports. The architecture consists of a *data management and preparation* layer, which performs document parsing, metadata enrichment, and extraction and normalization of textual and tabular content, and an *analytics layer*, which enables exploration, querying, and numerical reasoning over the extracted data. By bridging the gap between general-purpose LLM capabilities and the analytical requirements of ESG reporting, CLARIESG provides structured access to unstructured reports and supports accurate, explainable, and large-scale ESG analysis within a realistic regulatory context.

The demo illustrates two main scenarios. In the first, *ESG analysts* can query the system in natural language to obtain company and sector-level insights, with results returned both as explanatory text and as structured *scorecards* for benchmarking and decision-making. The second scenario focuses on *claim verification*, addressing the growing concern of *greenwashing* (Nemes et al., 2022; Moodaley and Telukdarie, 2023; de Freitas Netto et al., 2020), i.e., the practice of making unsubstantiated or misleading claims about a company's environmental performance. In this setting, *regulators, or sustainability promoters* can input a textual claim, and CLARIESG automatically retrieves and analyzes relevant evidence from sustainability reports, highlights supporting or contradicting passages, and provides natural-language justifications.

These scenarios demonstrate how CLARIESG supports analytical and regulatory needs, enhancing the reliability of ESG reporting.

2 Related Work

Recent approaches have explored LLM-based analysis of corporate and environmental reports. ESGReveal (Zou et al., 2023) leverages LLM reasoning to extract ESG indicators from both textual and tabular content, ChatReport (Ni et al., 2023) focuses on extracting traceable insights from sustainability reports while reducing hallucinations. ChatClimate (Vaghefi et al., 2023) integrates general-purpose LLM knowledge with climate-specific content, summarizing and distilling information relevant to user queries. FinQAPT (Singh et al., 2024) addresses QA over financial reports by combining dense retrieval, re-ranking, and LLM reasoning with dynamic n-shot prompting and chain-of-thought. NetZeroFacts (Wrzalik et al., 2024) extracts structured emissions data, enabling systematic analysis of corporate climate commitments.

Beyond report-centric approaches, some systems integrate further unstructured data sources. MyClimateCopilot (Nguyen et al., 2025) adopts an agentic framework that plans information retrieval, selects tools, and queries heterogeneous sources such as climate APIs and scientific literature. ChatNetZero (Hsu et al., 2024) similarly combines semantic retrieval with anti-hallucination strategies to extract answers from text and spreadsheets.

Another line of research focuses on optimizing individual components rather than providing end-to-end solutions. This includes approaches for extracting semantically structured ESG-related information from sustainability reports using LLMs (Zhou and Perzylo, 2023; Usmanova and Usbeck, 2024; Bronzini et al., 2024), as well as tools for parsing of reports with complex layouts, such as ReportParse (Morio et al., 2024).

Finally, several models have been fine-tuned on corporate and environmental reports, resulting in specialized architectures designed to address a variety of tasks, such as text classification (Xia et al., 2024; Mehra et al., 2022; Schimanski et al., 2023; Webersinke et al., 2021; Araci, 2019; Luukkonen et al., 2023), question answering (Luccioni et al., 2020; Zhao et al., 2022; Xie et al., 2023; Chen et al., 2021; Zhu et al., 2021; Deng et al., 2022; Wu et al., 2025), and claim extraction for greenwashing detection (Mahdavi et al., 2024).

Among the end-to-end systems, ESGReveal is the only approach that explicitly handles tabular structure, whereas FinQAPT uniquely focuses on complex numerical reasoning to effectively address quantitative queries. CLARIESG attempts to integrate both capabilities and resembles UNITQA (Zhu et al., 2025) in multi-table management, though the latter does not handle table extraction and retrieval from unstructured documents, instead assuming that tables are already available in relational or non-relational sources.

3 Approach

The architecture of CLARIESG is organized into two main layers, a data management and preparation layer and an analytics layer, as illustrated in Figure 1.

The *data management and preparation layer* handles the acquisition, parsing, and normalization of sustainability reports. Specifically, this layer integrates document-level parsing, company and sec-

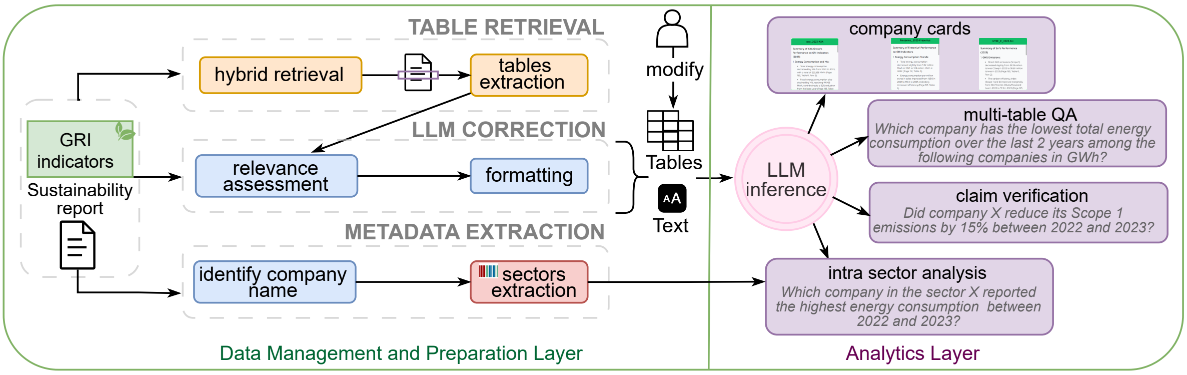


TABLE RETRIEVAL
hybrid retrieval
tables extraction
modify
company cards
LLM CORRECTION
GRI indicators
Sustainability report
relevance assessment
formatting
Tables
Text
LLM inference
multi-table QA
Which company has the lowest total energy consumption over the last 2 years among the following companies in GWh?
claim verification
Did company X reduce its Scope 1 emissions by 15% between 2022 and 2023?
METADATA EXTRACTION
identify company name
sectors extraction
intra sector analysis
Which company in the sector X reported the highest energy consumption between 2022 and 2023?
Data Management and Preparation Layer
Analytics Layer

Figure 1: System overview (🟦 use of LLM)

tor metadata retrieval, and the extraction and standardization of tabular and textual content according to ESG reporting frameworks such as GRI (GRI, 2024). Importantly, it adopts a human-in-the-loop approach (Amershi et al., 2014; Wu et al., 2022): while the process is largely automated, users can supervise the collected data, inspect intermediate results, and correct possible import or formatting errors. This optional supervision ensures reliability and traceability without compromising scalability. By producing consistent, validated representations, this layer lays the foundation for reproducible and large-scale ESG analysis.

The *analytics layer* builds upon these representations to enable interactive reasoning and knowledge discovery. It supports both natural-language querying and analytical operations such as numerical comparison, aggregation, and ranking across multiple documents. To this end, it combines LLM-based text understanding with reasoning paradigms such as Chain-of-Thought (Wei et al., 2022; Kojima et al., 2022) and Program-of-Thought (Chen et al., 2023).

By explicitly separating document understanding from reasoning and exploration, CLARIESG offers a unified framework for scalable, transparent, and explainable ESG analysis, anticipating the data landscape that will emerge with the introduction of the *European Single Access Point* (CSR, 2022).

3.1 Data Management and Preparation Layer

The data management and preparation layer is designed to transform raw sustainability reports into structured, validated, and searchable data representations. Its pipeline covers four main responsibilities: (*i*) identifying the target company and its metadata, (*ii*) locating the portions of the document that contain relevant information, (*iii*) extracting and validating tables and contextual text, and (*iv*) storing the resulting representations for subsequent retrieval and reasoning.

First, the system analyzes each report to identify the legal company name, which is then used to query Wikidata for sectoral metadata. Second, a hybrid sparse–dense[2] retrieval stage indexes the textual content in a vector database and localizes the sections of the document relevant to GRI topics, focusing the analysis on pages most likely to contain quantitative disclosures. Third, CLARIESG extracts and validates tabular data, which represent a large portion of ESG indicators. Tables in these reports are highly heterogeneous, often including multi-level headers, nested indicators, subtotals, or irregularly merged cells. An ensemble of OCR systems, combining `Unstructured`[3] and `Tesseract` (Smith, 2007), is used to extract table content. The extracted tables then undergo a two-step LLM-based refinement: a relevance assessment with respect to GRI indicators, followed by structural and formatting correction. Residual inaccuracies can be reviewed and corrected by the user through a human-in-the-loop interface, which allows inspection and validation against the original document. Finally, both refined tables and associated textual contexts are stored inside the database. Additional details on the prompts used for metadata extraction and table processing are provided in the Appendix (Section 7.1).

3.2 Analytics Layer

CLARIESG supports comparative and quantitative analyses that often require integrating information from multiple tables within a single report or across several reports. To address this challenge, CLARIESG implements a prompting-based workflow

[2] `multilingual-e5-large-instruct` and TF-IDF
[3] Unstructured Documentation

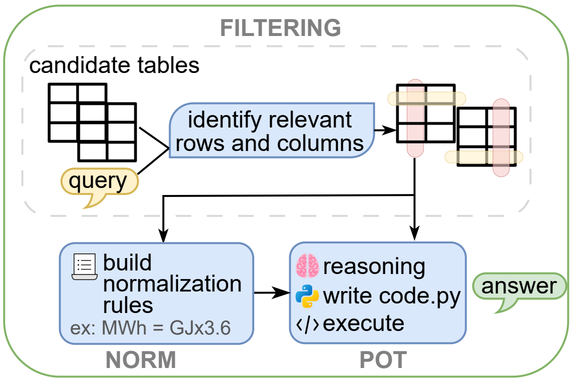


Figure 2: NormPoT prompting strategy. (🟦 use of LLM)

that orchestrates a sequence of LLM-guided reasoning steps for multi-table and quantitative tasks.

The process begins with *table filtering*, where the system analyzes candidate tables and selects the most relevant rows and columns according to the user query. This step is fully realized through prompting: the model is instructed to rank table segments by relevance to the question. The selected fragments are then passed to the *normalization* stage, where a second prompt alignes units, labels, and conventions to ensure consistency across heterogeneous sources. For example, energy consumption may appear as “GJ” in one report and “MWh” in another; ClariESG automatically applies conversion factors and aligns terminology to a standard schema, enabling direct comparison.

Next, *program generation* is performed using a Program-of-Thought (PoT) reasoning paradigm. The model is prompted to synthesize a Python function encoding the operations required to aggregate or combine data from the filtered and normalized tables. The generated code is parsed, sanitized, and executed locally to compute the requested metric. Finally, in the *answer composition* phase, the system reformulates the numeric output into a human-readable explanation that references both the computed value and the supporting evidence. We refer to this end-to-end reasoning pipeline as **NormPoT** (Figure 2). The Appendix (Section 7.2) provides the full set of prompts that operationalize each stage of the workflow.

3.3 Implementation details

ClariESG is implemented in Python. Data is stored in a PostgreSQL database extended with the `pgvector` module. `gpt-4o-mini` is used as a language model for structured information extraction and reasoning tasks. ClariESG is independent of the LLM, allowing the underlying model to be replaced, e.g. with open source alternatives, without modifying ClariESG’s pipeline. The user interface is built with Gradio 5.46.0, providing an interactive environment for uploading reports, inspecting tables, and executing queries. Interactions with the LLM are handled through the OpenAI API.

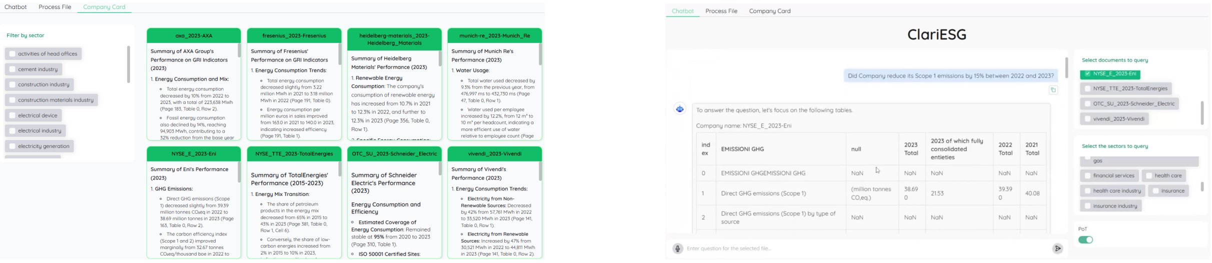
4 Application scenarios

This section illustrates two representative scenarios in which ClariESG supports automated analysis of corporate sustainability disclosures: (i) comparative ESG benchmarking, and (ii) claim verification.

4.1 Comparative ESG Analysis

This functionality represents a core step in several downstream applications, including the identification of top-performing and under-performing peers within a sector and the support of investment and policy decisions grounded in comparable evidence. Traditionally, analysts extract KPIs from individual sustainability reports, manually transfer them into spreadsheets, and attempt to harmonize units, time frames, and reporting methodologies. This harmonization process is tedious and prone to inconsistencies, particularly when reports differ in structure, terminology, or indicator granularity.

ClariESG automates this entire workflow by combining standard reasoning with comparison operations specifically designed for multi-table settings, such as value normalization and table-structure alignment. More precisely, ClariESG can extract, standardize, and interconnect corresponding indicators from multiple reports. The resulting processed information can be accessed through two complementary modalities: **company scorecards** and **conversational analytical responses**. In the first modality, the system transforms the semi-structured content of sustainability reports into interlinked, machine-readable data aligned with common reporting frameworks, such as the GRI standards. This data provides concise representations of relevant indicators. An example of output produced is shown in Figure 3a. In the second modality, users can interact directly with the system via a conversational interface. For instance, a query such as *“Which company reported the highest energy consumption in the manufacturing sector between 2022 and 2023?”* triggers the retrieval of the relevant values from each report, exe-



axa_2023-AXA

Summary of AXA Group's Performance on GRI Indicators (2023)

1. Energy Consumption and Mix:
   - Total energy consumption decreased by 10% from 2022 to 2023, with a total of 223,638 MWh (Page 183, Table 0, Row 2).
   - Fossil energy consumption also declined by 14%, reaching 94,903 MWh, contributing to a 32% reduction from the base year

fresenius_2023-Fresenius

Summary of Fresenius' Performance on GRI Indicators

1. Energy Consumption Trends:
   - Total energy consumption decreased slightly from 3.22 million MWh in 2021 to 3.18 million MWh in 2022 (Page 191, Table 0).
   - Energy consumption per million euros in sales improved from 163.0 in 2021 to 140.0 in 2023, indicating increased efficiency (Page 191, Table 1).

heidelberg-materials_2023-Heidelberg_Materials

Summary of Heidelberg Materials' Performance (2023)

1. Renewable Energy Consumption: The company's consumption of renewable energy has increased from 10.7% in 2021 to 12.3% in 2022, and further to 12.3% in 2023 (Page 356, Table 0, Row 1).

munich-re_2023-Munich_Re

Summary of Munich Re's Performance (2023)

1. Water Usage:
   - Total water used decreased by 9.3% from the previous year, from 476,997 ms to 432,730 ms (Page 47, Table 0, Row 1).
   - Water used per employee increased by 12.2%, from 12 m³ to 10 m³ per headcount, indicating a more efficient use of water relative to employee count (Page

NYSE_E_2023-Eni

Summary of Eni's Performance (2023)

1. GHG Emissions:
   - Direct GHG emissions (Scope 1) decreased slightly from 39.39 million tonnes CO₂eq in 2022 to 38.69 million tonnes in 2023 (Page 163, Table 0, Row 2).
   - The carbon efficiency index (Scope 1 and 2) improved marginally from 32.67 tonnes CO₂eq/thousand boe in 2022 to

NYSE_TTE_2023-TotalEnergies

Summary of TotalEnergies' Performance (2015-2023)

1. Energy Mix Transition:
   - The share of petroleum products in the energy mix decreased from 65% in 2015 to 43% in 2023 (Page 381, Table 0, Row 1, Cell 6).
   - Conversely, the share of low-carbon energies increased from 2% in 2015 to 10% in 2023,

OTC_SU_2023-Schneider_Electric

Summary of Schneider Electric's Performance (2023)

Energy Consumption and Efficiency

- **Estimated Coverage of Energy Consumption**: Remained stable at **95%** from 2020 to 2023 (Page 310, Table 1).
- **ISO 50001 Certified Sites**:

vivendi_2023-Vivendi

Summary of Vivendi's Performance (2023)

1. Energy Consumption Trends:
   - **Electricity from Non-Renewable Sources:** Decreased by 42% from 57,761 MWh in 2022 to 33,520 MWh in 2023 (Page 141, Table 0, Row 1).
   - **Electricity from Renewable Sources:** Increased by 47% from 30,521 MWh in 2022 to 44,811 MWh in 2023 (Page 141, Table 0, Row 2).

(a) Company scorecards.

Did Company reduce its Scope 1 emissions by 15% between 2022 and 2023?

To answer the question, let's focus on the following tables.

Company name: NYSE_E_2023-Eni

index	EMISSIONI GHG	null	2023 Total	2023 of which fully consolidated entieties	2022 Total	2021 Total
0	EMISSIONI GHGEMISSIONI GHG	NaN	NaN	NaN	NaN	NaN
1	Direct GHG emissions (Scope 1)	(million tonnes CO,eq.)	38.690	21.53	39.390	40.08
2	Direct GHG emissions (Scope 1) by type of source	NaN	NaN	NaN	NaN	NaN

(b) Conversational interface.

Figure 3: Key CLARIESG functionalities: automated ESG comparison and interactive claim verification.

cution of the necessary computations, and formulation of a unified answer with references to the original sources. This capability enables reproducible, and transparent quantitative ESG benchmarking, thus supporting interactive and data-driven exploration of corporate sustainability at scale.

4.2 Claim verification

The second application domain concerns the verification of claims and the answering of questions within corporate sustainability reports (see the conversation interface in Figure 3b). This task is central to the identification of potential greenwashing, as it enables the detection of inconsistencies between narrative claims and reported evidence. In traditional workflows, verifying such claims requires labor-intensive manual inspection of lengthy reports and the cross-referencing of textual statements with tabular indicators dispersed across multiple sections. This procedure is inherently slow, error-prone, and difficult to scale, particularly when dealing with heterogeneous reporting formats or multiple companies. In contrast, CLARIESG automates the task and enhances its effectiveness. By leveraging numerical reasoning capabilities through a Program-of-Thought paradigm, the system allows users to query ESG disclosures in natural language with questions such as *"What is the percentage reduction of greenhouse gas emissions since 2020?"*. The model identifies the main indicators and temporal references in textual and tabular evidence, computes the required numerical variation, and returns an evidence-grounded answer that explicitly includes the supporting excerpts and table cells. This design enables analysts to audit the full reasoning chain.

5 Main results

A core requirement in the proposed scenarios is accurately resolving QA tasks over single and multiple tables. To evaluate this, we assess CLARIESG on GRI-QA, a domain-specific QA benchmark over environmental tables from sustainability reports, which allows us to replicate the core operation behind both scenarios in a controlled setting. While GRI-QA specifically targets GRI 300 indicators, CLARIESG is GRI-agnostic and can be extended to other GRI families by updating indicator metadata. GRI-QA organizes the questions into the following categories: *extractive* questions that require direct data retrieval; *hierarchical* questions that involve disambiguating terms within nested table structures; and *calculated* and *quantitative* questions that test relational and arithmetic reasoning such as comparisons, superlatives, rankings, and percentage variations. It also includes *multi-step* questions requiring computations over multiple tables or documents.

By analyzing CLARIESG 's responses on GRI-QA, we assess (i) the effectiveness of prompting strategies for *single-table* and *multi-table* questions, and (ii) the relative performance of ChatGPT 5.1 and CLARIESG in multi-table reasoning. We use the normalized Exact Match (EM) (Dua et al., 2019) as the main evaluation metric.

Comparison of Prompting Strategies. Table 1 shows that the simple use of Chain-of-Thought (CoT) on *one-table* questions provides the best performance. In particular, the average performance of Program-of-Thoughts (PoT) and NormPoT decreases by 7 and 7.9 EM points respectively, demonstrating that overly complex prompting strategies can lead to sub-optimal performance on simple questions. The only *one-table* scenario where the performance of PoT and NormPoT improves compared to CoT is for the quant dataset, where performance increases by 13.1 and 13.5 points respectively. This indicates that for questions requiring mathematical calculations, performing the calculations through a Python interpreter leads to better results. For the *multi-table*

	GRI-QA one-table						*GRI-QA multi-table*									
	extra	hier	rel	quant	step	avg	rel2	rel3	rel5	quant2	quant3	quant5	step2	step3	step5	avg
CoT	84.2	80.9	92.7	72.6	33.1	72.7	56.6	34.3	19.5	58.7	20.8	0.0	43.7	32.7	25.5	32.4
PoT	62.4	66.8	89.9	85.7	23.5	$65.7^{-7.0}$	63.0	41.0	26.4	65.3	36.1	12.0	58.9	37.0	30.9	$41.2^{+8.8}$
NormPoT	63.9	62.2	91.5	86.1	20.5	$64.8^{-7.9}$	68.5	59.0	39.1	69.3	50.0	22.0	56.3	42.8	28.2	$48.4^{+16.0}$

Table 1: Performance of different prompting strategies in one-table and multi-table settings, and for the question categories defined in GRI-QA. In multi-table tasks, the number beside each category indicates the tables involved (e.g., rel5 = 5 tables). Superscripts denote average performance differences relative to the CoT baseline.

		rel2	rel3	rel5	quant2	quant3	quant5	step2	step3	step5	avg
	ChatGPT 5.1	70.0	50.0	30.0	50.0	16.0	28.0	60.0	40.0	36.0	42.2
ClariESG	gpt-4o-mini	60.0	56.0	44.0	72.0	46.0	22.0	72.0	44.0	26.0	$49.1^{+6.9}$
	gpt-4o-mini + noisy tables	44.0	28.0	30.0	68.0	28.0	0.0	54.0	40.0	30.0	$35.8^{-6.4}$
	gpt-5-mini	90.0	88.0	88.0	92.0	74.0	68.0	76.0	70.0	66.0	$79.1^{+36.9}$
	gpt-5-mini + noisy tables	88.0	90.0	84.0	88.0	74.0	66.0	76.0	72.0	60.0	$77.6^{+35.4}$

Table 2: Performance comparison between ChatGPT 5.1 and ClariESG (with different LLMs) on the first 50 questions of each GRI-QA *multi-table* benchmark. In the noisy setting, two irrelevant tables are added for each company report. Superscripts indicate average performance deltas vs. ChatGPT 5.1.

benchmarks, on the other hand, PoT and NormPoT achieve average performance that is 8.8 and 16 EM points higher than CoT, respectively. In particular, the integration of a normalization step prior to executing PoT provides significant advantages, by clarifying the intermediate steps required to compare values from different companies that use different units of measurement. The results indicate that for questions requiring numerical calculation or reasoning across multiple tables, the best strategy to adopt is NormPoT.

Comparison with ChatGPT 5.1. To validate the quality of ClariESG, we compare its performance on the first 50 questions of each *multi-table* benchmark of GRI-QA with ChatGPT 5.1. The systems are compared based on how they would be used to perform ESG analysis of corporate reports. For ChatGPT 5.1, for each question, we manually connect to the ChatGPT website, we load the complete reports of the companies required by the question and we annotate its response. Based on the request, ChatGPT 5.1 itself decides whether to think longer (*ChatGPT Thinking*) or provide an immediate answer (*ChatGPT Instant*). For ClariESG, we use the clean tables provided by GRI-QA. Although the amount of textual context provided as input differs between ChatGPT 5.1 and ClariESG, the comparison between the two systems is fair, assuming that the user correctly cleans the tables extracted by ClariESG. Still, to faithfully evaluate the performance of ClariESG, we also test it with two additional *noisy* tables as context for each company needed to answer the question.

Table 2 shows the results. In ClariESG, the average performance of gpt-4o-mini surpasses the performance of ChatGPT 5.1 by 6.9 EM points when ClariESG is not provided with additional *noisy* tables, whereas its performance falls 6.4 EM points below that of ChatGPT 5.1 when evaluated under the *noisy* setting. By using gpt-5-mini as backbone LLM, ClariESG greatly outperforms ChatGPT 5.1 with a respective average performance increase of 36.9 and 35.4 EM points for the *clean* and *noisy* settings. Notably, providing clean tables and correct context allows ClariESG to outperform ChatGPT 5.1 with both gpt-4o-mini and gpt-5-mini, even if the backbone LLM is much smaller. In general, the tabular data management and reduction of context performed by ClariESG proves to be crucial in providing accurate responses to ESG-related queries.

6 Conclusion

We showcased ClariESG, an end-to-end system for analyzing corporate sustainability reports. Combining robust table extraction with structured prompting for multi-table normalization and Program-of-Thoughts reasoning, ClariESG provides precise, auditable analytics for ESG benchmarking and claim verification. Experiments on

GRI-QA show that this specialized workflow outperforms general-purpose LLMs such as ChatGPT 5.1. Future work will focus on improving robustness to noisy tables and integrating richer domain knowledge.

Limitations

The system currently focuses exclusively on the management and extraction of information related to ESG data. This represents an essential step for enabling analysts to gain a deeper understanding of companies' environmental behaviour and to compare performance across sectors. However, integrating indicators that combine ESG and financial data would further enhance the analytical value of the system, as investment decisions are often guided by a combination of both dimensions. We plan to address this limitation in future work.

OpenAI models accessed via API calls are known to produce non-deterministic outputs even when the temperature is set to 0. As a result, the results reported in Table 1 and Table 2 may exhibit slight variability across different runs.

Risks

A potential risk associated with the use of CLARIESG is that analysts may over-rely on the system's responses. Although the performance of CLARIESG is promising (Table 2), it is not flawless. Even though the reasoning process used to generate answers is fully auditable, users may still place trust in the output without verifying the underlying evidence. For this reason, we recommend that analysts consult CLARIESG as a support tool, but cross-check its answers against the original sources to prevent misinterpretations and mitigate the possibility of hallucinations.

Use of AI Assistants

When writing this paper, we used AI assistants, such as ChatGPT, to improve the flow of writing and the vocabulary of the initial drafts we manually wrote. Each suggestion has been manually validated by the authors.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support. Moreover, this work was partially funded by the RESIST0 project (PR-FESR Emilia-Romagna 2021-2027) through a grant to the AIRI research center at the University of Modena and Reggio Emilia.

References

2022. Directive (eu) 2022/2464 of the european parliament and of the council of 14 december 2022 on corporate sustainability reporting (csrd). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2464. Official Journal of the European Union, L 322, 16 December 2022.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, pages 483–490. AAAI Press.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063.

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 Sci.*, 13(1):41.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Vienna, Austria. Association for Computational Linguistics.

Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. 2020. Concepts and forms of greenwashing: a systematic review. *Environmental Sciences Europe*, 32(1):19.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. PACIFIC: towards proactive conversational question answering over tabular and textual data in finance. In *EMNLP*, pages 6970–6984. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

GRI. 2024. Global reporting initiative website.

A. Heflich and J. Saulnier. 2024. Potential economic impact of european sustainable finance. Technical report, European Parliamentary Research Service.

Angel Hsu, Mason Laney, Ji Zhang, Diego Manya, and Linda Farczadi. 2024. Evaluating ChatNetZero, an LLM-chatbot to demystify climate pledges. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 82–92, Bangkok, Thailand. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. Analyzing sustainability reports using natural language processing. *CoRR*, abs/2011.08073.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, and 2 others. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

Mohammad Mahdavi, Ramin Baghaei Mehr, and Tom Debus. 2024. Combat greenwashing with goalspotter: Automatic sustainability objective detection in heterogeneous reports. In *CIKM*, pages 4752–4759. ACM.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. ESGBERT: language model to help with classification tasks related to companies environmental, social, and governance practices. *CoRR*, abs/2203.16788.

Wayne Moodaley and Arnesh Telukdarie. 2023. Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review. *Sustainability*, 15(2).

Gaku Morio, Soh Young In, Jungah Yoon, Harri Rowlands, and Christopher D. Manning. 2024. Reportparse: A unified NLP tool for extracting document structure and semantics of corporate sustainability reporting. In *IJCAI*, pages 8749–8753. ijcai.org.

Noémi Nemes, Stephen J. Scanlan, Pete Smith, Tone Smith, Melissa Aronczyk, Stephanie Hill, Simon L. Lewis, A. Wren Montgomery, Francesco N. Tubiello, and Doreen Stabinsky. 2022. An integrated framework to assess greenwashing. *Sustainability*, 14(8).

Vincent Nguyen, Willow Hallgren, Ashley Harkin, Mahesh Prakash, and Sarvnaz Karimi. 2025. My climate CoPilot: A question answering system for climate adaptation in agriculture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 62–70, Vienna, Austria. Association for Computational Linguistics.

Jingwei Ni, Julia Anna Bingler, Chiara Colesanti Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: democratizing sustainability disclosure analysis through llm-based tools. In *EMNLP (Demos)*, pages 21–51. Association for Computational Linguistics.

Tobias Schimanski, Julia Anna Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. Climatebert-netzero: Detecting and assessing net zero and reduction targets. In *EMNLP*, pages 15745–15756. Association for Computational Linguistics.

Kuldeep Singh, Simerjot Kaur, and Charese Smiley. 2024. Finqapt: Empowering financial decisions with end-to-end llm-driven question answering pipeline. In *ICAIF*, pages 266–273. ACM.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

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, Bangkok, Thailand. Association for Computational Linguistics.

Saeid Ashraf Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Anna Bingler, Tobias Schimanski, Chiara Colesanti Senni, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. chatclimate: Grounding conversational AI in climate science. *CoRR*, abs/2304.05510.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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, Torino, Italia. Association for Computational Linguistics.

Tianyi Wu, Wei Fan, Junjie Wu, and Hui Xiong. 2022. A survey on human-in-the-loop machine learning: Challenges and opportunities. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–31.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Tongliang Li, Zhoujun Li, and Guanglin Niu. 2025. Tablebench: A comprehensive and complex benchmark for table question answering. In *AAAI*, pages 25497–25506. AAAI Press.

Lei Xia, Mingming Yang, and Qi Liu. 2024. Using pretrained language model for accurate ESG prediction. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 1–22, Jeju, South Korea. -.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. In *Advances in Neural Information Processing Systems*, volume 36, pages 33469–33484. Curran Associates, Inc.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *ACL (1)*, pages 6588–6600. Association for Computational Linguistics.

Yuchen Zhou and Alexander Perzylo. 2023. Ontosustain: Towards an ontology for corporate sustainability reporting. In *ISWC (Posters/Demos/Industry)*, volume 3632 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

Jun-Peng Zhu, Peng Cai, Kai Xu, Li Li, Yishen Sun, Shuai Zhou, Haihuang Su, Liu Tang, and Qi Liu. 2025. UNITQA: A unified automated tabular question answering system with multi-agent large language models. In *SIGMOD Conference Companion*, pages 279–282. ACM.

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. *CoRR*, abs/2312.17264.

7 Prompts and screenshots

The screenshot in Figure 4 shows the component used to upload and refine the tables extracted from the reports, which could not be included in the main paper due to space constraints.

Below, instead, we provide details on the prompts used to instruct the underlying LLM of CLARIESG to perform the different tasks required by the system. We distinguish between prompts employed for data preparation and those used for analysing the extracted data.

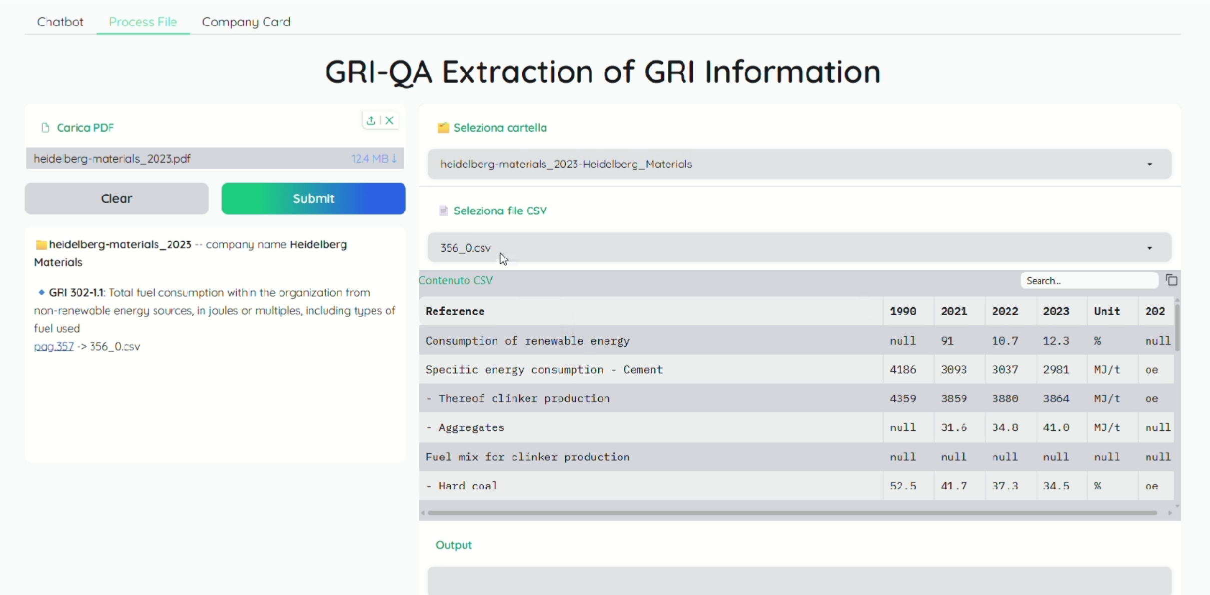
7.1 Prompts for the data management and preparation layer

During the pre-processing phase, the LLM is responsible for (i) extracting metadata about the reporting company (such as the legal name and industrial sector), and (ii) accurately identifying the tables contained in the document. Specifically, Figure 5 shows the prompt used to identify the company's legal name from the front pages of the report. The extracted name is then used to query **Wikidata** and retrieve the company's industrial sectors. The SPARQL query used for this retrieval is provided in Figure 6. To ensure robust table extraction, OCR output is further processed through a two-step LLM-based pipeline. This includes (i) verifying the relevance of the extracted content with respect to GRI indicators (see the prompt in Figure 7), and (ii) refining the structural and formatting consistency of the resulting tables (see the prompt in Figure 8).

7.2 Prompts for the analytics layer

To support comparative analysis over tables extracted from multiple reports, CLARIESG orchestrates a sequence of LLM-guided reasoning steps. These include: (i) *table filtering*, to select rows and columns relevant to the user query (see the prompt in Figure 9); (ii) *normalization*, to harmonize units, labels, and formatting conventions across heterogeneous sources (see the prompt in Figure 10); (iii) *program generation*, which synthesizes a Python function encoding the required logical and arithmetic operations in a PoT-style workflow (see the prompt in Figure 12); and (iv) *code execution*, to run the generated program and obtain the final computed values. In case of Python execution error, CLARIESG falls back to standard CoT (see the prompt in Figure 11).

Additionally, CLARIESG uses the prompt in Figure 13 to create the *company scorecards*.



GRI-QA Extraction of GRI Information

Carica PDF

heidelberg-materials_2023.pdf 12.4 MB

heidelberg-materials_2023 -- company name **Heidelberg Materials**

◆ **GRI 302-1.1**: Total fuel consumption within the organization from non-renewable energy sources, in joules or multiples, including types of fuel used
pag.357 -> 356_0.csv

Seleziona cartella

heidelberg-materials_2023-Heidelberg_Materials

Seleziona file CSV

356_0.csv

Contenuto CSV

Reference	1990	2021	2022	2023	Unit	202
Consumption of renewable energy	null	91	10.7	12.3	%	null
Specific energy consumption - Cement	4186	3093	3037	2981	MJ/t	oe
- Thereof clinker production	4359	3859	3880	3864	MJ/t	oe
- Aggregates	null	31.6	34.8	41.0	MJ/t	null
Fuel mix for clinker production	null	null	null	null	null	null
- Hard coal	52.5	41.7	37.3	34.5	%	oe

Output

Figure 4: Screen of Tab 2, Upload and processing documents.

Get company name

You are an assistant that extracts the name of the main company mentioned in a PDF document.

Read the following text and return ONLY the company's full name — no explanations, no punctuation, no additional text.

Figure 5: Prompt to obtain the company name.

Get company sectors

```
SELECT ?company ?companyLabel ?industry ?industryLabel WHERE {{
  SERVICE wikibase:mwapi {{
     bd:serviceParam wikibase:endpoint "www.wikidata.org";
         wikibase:api "EntitySearch";
         mwapi:search "company_name";
         mwapi:language "en".
     ?company wikibase:apiOutputItem mwapi:item .
}}
  OPTIONAL {{ ?company wdt:P452 ?industry.}}
  SERVICE wikibase:label{{ bd:serviceParam wikibase:language "\[AUTO_LANGUAGE],en".}}
}}
LIMIT 10
```

Figure 6: Wikidata query to retrieve the industry sectors associated with a company.

Evaluation of extracted tables

You are an expert in sustainability reporting (GRI Standards).

I will give you:
1. A GRI code and its description.
2. The content of a CSV table extracted from a company report.

Task: Decide if this CSV table is relevant to the GRI code.

Answer with ONLY one word: "YES" if the CSV contains information that matches or supports the GRI description, otherwise "NO".

GRI code: gri_code
Description: gri_desc
CSV content (partial preview): csv_preview

Figure 7: Prompt to evaluate the extracted tables.

Formatted table

You are given the content of a CSV file automatically extracted from a table.
Your task is to clean and reformat it into a valid table, ensuring that **all rows have the same number of columns**.

Follow these rules strictly:

- Use **;** as the column separator in the final output.
- Determine the **maximum number of fields** present in any row, and expand all rows to that length.
- If a row has missing cells, fill them with NaN.
- Keep numeric values as-is, including negative percentages and decimals.
- Fix broken or merged cells, misplaced values, or incorrect headers.
- **Do not add or remove data rows** except for lines that are completely empty or contain only NaN.
- Standardize headers:
  - Create clear, readable names.
  - Avoid duplicates (rename automatically if needed).
  - Do not lose or shorten the meaning of headers.
- Ensure consistent formatting:
  - Align numeric and text values properly.
  - Remove symbols or characters that are clearly OCR or extraction noise.
- Output **only the cleaned CSV content** no explanations or comments.

REMEMBER THAT ALL ROWS MUST HAVE THE SAME NUMBER OF FIELDS!

Figure 8: Prompt to format the extracted tables.

Extract values

You will be given a question and a table.
ONLY IF there are relevant rows and columns, you must indicate the indices of the rows and columns that could be relevant to answer the question. OTHERWISE, if for a certain table there are no relevant rows and columns, write an empty list for both "rows" and "columns". You must not try to answer the question, you must only retrieve the relevant rows if there are. Use the values in the "index" column to refer to the relevant rows.
Additionally, for each selected row include the corresponding row name in the table: use the value from the first non-index column (immediately to the right of "index") as the row's name. Align "row_names" with "rows". If no such column exists, use an empty string (”).

For column indices, write the number (starting from 0, left to right), not the column name. First reason step-by-step. Then write "Final answer: " followed exclusively by a Python dictionary:

```
{
    "rows": [row_index1,...,row_indexn],
    "columns": [column_index1,...,column_indexn],
    "row_names": [row_name1,...,row_namen]
}
```

If no relevant rows/columns, return empty lists. Do not write anything else after "Final answer:". Do not use Markdown syntax.

Question: {question}
Table: {table}

Let's think step-by-step.

Figure 9: Prompt to extract relevant rows and columns from a table.

Normalization

Given multiple tables and a question, decide the unit of measure to use for the final answer. Then, align table values by converting needed values to a unique unit.

If the question specifies a unit, convert values to it. Otherwise, decide the unit and convert. Do not rewrite the tables. Only provide a list of rules/formulas indicating the needed transformations. Transformations must only handle units. Do not discuss solving the question.

Sample rule: 1. 1000 meters = 1 kilometer

First reason step-by-step. Then write "Final answer: " followed exclusively by the list of rules/formulas.

Do not write anything else after "Final answer:". Do not use Markdown syntax.

Question: {question}
Tables: {tables}

Let's think step-by-step.

Figure 10: Normalize units in multiple tables.

Fallback Python

Consider the following question and content. First reason step-by-step, then provide the answer.

Question: {question}
Content: {content}

Let's think step-by-step.

Figure 11: Prompt to reason step-by-step and provide Python answer.

Python Prompt

You need to create Python code that answers the following question, taking into account the tables provided and the fact that NOT ALL rows are always useful for generating the answer. Write your reasoning first. Then, at the end, write 'Final answer:' followed by the Python code and nothing else. The Python code must be executable 'as is', so include relevant imports. At the end, print the result with print(). If not already done,specify ```python before the code and ``` at the end.

If the question is Boolean, the output must be exclusively 'yes' or 'no'. If a list of values is required, respond with a comma-separated list. Write numerical values with exactly 2 decimal places.

Ensure the final answer is in the expected form. Do not write anything else after 'Final answer:'. Do not use Markdown syntax.

Question: {question}
Tables: {paragraph}

Let's think step by step.

Figure 12: Prompt to generate Python code considering relevant rows/columns in tables.

Company Summary

You are an expert assistant in sustainability and GRI standards.

Your task is to analyze data extracted from a company's PDFs in the form of CSV tables related to specific GRI indicators, and provide a clear, concise summary of the company's performance.

Instructions:

- Base your summary strictly on the data provided in the CSV tables.
- Highlight trends, improvements, or regressions in the company's performance where possible.
- Do not add assumptions or information not present in the tables.
- For each key point, reference the row, cell, page, and table number used from the CSV context.
- Make the summary concise, well-structured, and readable for stakeholders.
- If there is no context, reply clearly that you have not received any information. Nothing else.

Here are the CSV tables extracted from the company's PDFs related to GRI indicators:

{context}

Please provide a concise summary of the company's performance based strictly on this data.

Figure 13: Prompt to generate a concise summary of company performance from GRI-related CSV tables as company card.